

# Asymptotic properties of parallel Bayesian kernel density estimators

Alexey Miroshnikov \*

Evgeny Savelev<sup>†</sup>

## Abstract

In this article we perform an asymptotic analysis of Bayesian parallel kernel density estimators introduced in [19]. We derive the asymptotic expansion of mean integrated square error for the full data posterior estimator and investigate the properties of asymptotically optimal bandwidth parameters. Our analysis demonstrates that partitioning data in subsets affects significantly the values of optimal parameters.

## 1 Introduction

Recent developments in data science and analytics research have produced an abundance of large data sets that are too large to be analyzed in their entirety. As the size of data sets increases, the time required for processing rises significantly. An effective solution to this problem is to perform statistical analysis of large data sets with the use of parallel computing. The prevalence of parallel processing of large data sets motivated a surge in research on parallel statistical algorithms.

One approach is to divide data sets into smaller subsets, and analyze the subsets on separate machines using parallel Markov chain Monte Carlo (MCMC) methods [14, 18, 24]. These methods, however, require communication between machines for generation of each sample. Communication costs in modern computer networks dwarf the speed up achieved by parallel processing and therefore algorithms that require extensive communications are not desirable.

To address these issues, numerous alternative communication-free parallel MCMC methods have been developed for Bayesian analysis of big data. These methods partition data into subsets, perform independent Bayesian MCMC analysis on each subset, and combine the subset posterior samples to estimate the full data posterior see [23, 19, 17].

Neiswanger, Wang and Xing [19] introduced a parallel kernel density estimator that first approximates each subset posterior density; the full data posterior is then estimated by multiplying the subset posterior estimators together,

$$\hat{p}(\mathbf{x}|\mathbf{y}) \propto \hat{p}^*(\mathbf{x}|\mathbf{y}) := \hat{p}_1(\mathbf{x}|\mathbf{y}_1) \cdot \hat{p}_2(\mathbf{x}|\mathbf{y}_2) \cdots \hat{p}_M(\mathbf{x}|\mathbf{y}_M). \quad (1.1)$$

Here  $\mathbf{x} \in \mathbb{R}^d$  is the model parameter,  $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_M\}$  is the full data set partitioned into  $M$  disjoint independent subsets  $\mathbf{y}_m$ , and

$$\hat{p}_m(\mathbf{x}|\mathbf{y}_m) = \sum_{i=1}^{N_m} \frac{1}{h_m} K\left(\frac{\mathbf{x} - \mathbf{X}_i^m}{h_m}\right)$$

is the subset posterior kernel density estimator, with  $h_m \in \mathbb{R}_+$  a kernel bandwidth parameter.

---

\*Department of Mathematics, University of California, Los Angeles, amiroshn@gmail.com

<sup>†</sup>Department of Mathematics, Virginia Polytechnic Institute and State University, savelev@vt.edu

The authors of [19] show that the estimator (1.1) is asymptotically exact and develop an algorithm that generates samples from the distribution approximating the full data estimator. Though the estimator is asymptotically exact the algorithm of [19] does not perform well for posteriors that have non-Gaussian shape and, in addition, it is very sensitive to the choice of the kernel bandwidth parameters.

In the present article, we are concerned with the asymptotic analysis of the parallel Bayesian kernel density estimators of the form (1.1). In particular, we are interested in the asymptotic analysis of the mean integrated square error for the likelihood  $\hat{p}^*$  and density estimator  $\hat{p}$  as well as the properties of the optimal kernel bandwidth vector  $\mathbf{h}$ .

The kernel density estimators for the case  $M = 1$  have been studied extensively in the past five decades. In particular, asymptotic properties of mean integrated square error for the estimator (1.1) with  $M = 1$  and  $d = 1$  were obtained by Rosenblatt [7], Parzen [9] and Epanechnikov [8]. In particular, for sufficiently smooth probability densities Parzen [9] derived the asymptotic expansion

$$\text{MISE}[p, \hat{p}, \mathbf{N}, \mathbf{h}] = \frac{h^4 k_2^2}{4} \int_{\mathbb{R}} (p''(x))^2 dx + \frac{1}{nh} \int_{\mathbb{R}} K^2(t) dt + o\left(\frac{1}{nh} + h^4\right) \quad (1.2)$$

and obtained a formula for the asymptotically optimal bandwidth parameter

$$h_{M=1}^{\text{opt}} = n^{-1/5} k_2^{-2/5} \left( \int_{\mathbb{R}} K^2(t) dt \right)^{1/5} \left( \int_{\mathbb{R}} (p''(x))^2 dx \right)^{-1/5}, \quad (1.3)$$

which minimizes the leading terms in the expansion.

The case of non-differentiable or discontinuous probability density functions has been shown to possess different asymptotic estimates for MISE. The optimal bandwidth parameter  $h_{M=1}^{\text{opt}} \in \mathbb{R}$  and the rate of convergence of MISE depend directly on the regularity of  $p$  (see van Eden [6]).

In the case of multivariate distributions,  $\mathbf{x} \in \mathbb{R}^d$ , the complexity of the asymptotic analysis depends on the form of the bandwidth matrix  $\mathbf{H} \in \mathbb{R}^{d \times d}$ . In the simplest case, one can assume that  $\mathbf{H} = h\mathbf{I}$ , where  $h$  is a scalar (see Silverman [25], Simonoff [26] and Epanechnikov [8]). Another approach is to consider the bandwidth matrix of the form  $\mathbf{H} = \text{diag}(h_1, h_2, \dots, h_d)$ , with  $h_i$  being a bandwidth parameter for each dimension  $i \in \{1, \dots, d\}$ . The most general formulation assumes that  $\mathbf{H}$  is a  $d \times d$  matrix, which allows one to encode correlations between components of  $\mathbf{x}$  (see Duong and Hazelton [5], and Wand and Jones [29]).

In our work, motivated by the ideas of [9, 5, 29, 7] we consider the asymptotic analysis of the mean integrated square error for both the parallel likelihood estimator

$$\text{MISE}[\hat{p}^*, p^*; \mathbf{N}, \mathbf{h}] = \mathbb{E} \int_{\mathbb{R}} \left\{ p^*(x|\mathbf{y}) - \hat{p}^*(x|\mathbf{y}) \right\}^2 dx$$

and the full data set posterior density estimator

$$\text{MISE}[\hat{p}, p; \mathbf{N}, \mathbf{h}] = \mathbb{E} \int_{\mathbb{R}} \left\{ p(x|\mathbf{y}) - \hat{p}(x|\mathbf{y}) \right\}^2 dx.$$

as

$$\mathbf{N} = (N_1, N_2, \dots, N_M) \rightarrow \infty, \quad \mathbf{h} = (h_1, h_2, \dots, h_M) \rightarrow 0 \quad \text{and} \quad (\mathbf{N} \cdot \mathbf{h})^{-1} \rightarrow 0.$$

Under appropriate condition on the regularity of the probability density we derive the formula (3.27) for the leading part of the MISE for the likelihood  $\hat{p}^*$ , called  $\text{AMISE}[p^*, \hat{p}^*]$ . The leading part turns out to be in agreement with the leading part for  $M = 1$ , but for  $M > 1$  it is different from the leading terms in the expansion (1.2), or the expansion in the case  $M = 1$  and  $d = M$ .

The asymptotically optimal bandwidth parameter for the likelihood is then defined to be a minimizer

$$\mathbf{h}_*^{\text{opt}} = \text{argmin}_{\mathbf{h} \in \mathbb{R}_+^M} \text{AMISE}[p^*, \hat{p}^*; \mathbf{N}, \mathbf{h}].$$

In Theorem 3.5 we prove that the minimizer exists and, in fact, under certain conditions on the regularity of the probability density  $p$ , is unique. For certain, special cases of posterior densities we obtain the exact expressions (3.33) and (3.35) for the optimal bandwidth parameters.

To carry out the same program for the mean square error of the full data set posterior density in general is not possible because the renormalization constant

$$\hat{c} = \left( \int \hat{p}_1(x|\mathbf{y}) \cdot \hat{p}_2(x|\mathbf{y}) \dots \hat{p}_M(x|\mathbf{y}) dx \right)^{-1} = \left( \int \hat{p}^*(x|\mathbf{y}) dx \right)^{-1}$$

may in general have an infinite expectation. This may happen because on some events the estimators  $\hat{p}_i^*$  may decay too quickly in  $x$  variable and the sets of  $x$  with the most ‘mass’ for each  $\hat{p}_i^*$  may have little common intersection which potentially may lead to large values of  $\hat{c}$ . To avoid this situation one would need to chose the kernel  $K$  in appropriate way and establish the finiteness of the expectation of  $\hat{c}$ . In this article we, however, do not investigate this. Instead, we show that one can replace the mean integrated square error by an asymptotically equivalent distance functional denoted

$$\overline{\text{MISE}}[\hat{p}, p; \mathbf{N}, \mathbf{h}].$$

is defined in 4.13. The new functional is shown to be well-defined on the whole sample space  $\Omega$  and it is asymptotically equivalent to  $\text{MISE}(p, \hat{p})$  restricted to smaller events whose probability tends to one.

In our work we analyze the functional  $\overline{\text{MISE}}$  by carrying out the same program as for the likelihood. We first derive the leading part of the  $\overline{\text{MISE}}$ , given in (4.17), for the full data set posterior density  $\hat{p}$ , called  $\overline{\text{AMISE}}[p^*, \hat{p}^*]$ . The asymptotically optimal bandwidth parameter for the full data set posterior is then defined to be a minimizer

$$\mathbf{h}^{\text{opt}} = \operatorname{argmin}_{\mathbf{h} \in \mathbb{R}_+^M} \overline{\text{AMISE}}[p^*, \hat{p}^*; \mathbf{N}, \mathbf{h}].$$

We show that the minimizer exists and under certain conditions it is unique. In addition, for certain cases of posterior densities we obtain exact formula (4.19) for the optimal bandwidth vector.

Our analysis demonstrates that partitioning data into  $M > 1$  sets affects the optimality condition of parameter  $\mathbf{h}$ . It also indicates that the bandwidth vector

$$\mathbf{h}_0^{\text{opt}} = (h_{1,M=1}^{\text{opt}}, h_{2,M=1}^{\text{opt}}, \dots, h_{M,M=1}^{\text{opt}})$$

where  $h_{m,M=1}^{\text{opt}}$  is an optimal bandwidth parameter for the estimator  $\hat{p}_m(x|\mathbf{y})$  computed by (1.3), is suboptimal for both estimators  $\hat{p}^*$  and  $\hat{p}$  whenever  $M > 1$ . This observation highlights the fact that the choice of optimal parameters for parallel kernel density estimators (suitable for parallelizing data analysis) must differ from the theoretical choice suggested in case of processing on a single machine.

The structure of the article is as follows. In Section 2 we introduce the notation used throughout the article and hypotheses that form the foundation of the analysis. In Section 3 we perform the asymptotic analysis of  $\text{MISE}$  for the likelihood. In Section 4 we perform the analysis of  $\text{MISE}$  for the full data set posterior density. Section 5 is an appendix that contains lemmas and theorems we employ to obtain the main results of Section 3 and Section 4.

## 2 Notation and hypotheses

For the convenience of the reader we collect in this section all hypotheses and results relevant to our analysis and present the notation that is utilized throughout the article.

**(H1)** Motivated by the form of the posterior density at Neiswanger et al. [19] we consider the probability density function of the form

$$p(x) \propto p^*(x) \quad \text{where} \quad p^*(x) := \prod_{m=1}^M p_m(x) \quad (2.1)$$

**(H2)** For each  $m \in \{1, \dots, M\}$   $p_m(x)$  is a probability density function. We consider the estimator of  $p$  in the form

$$\hat{p}(x) \propto \hat{p}^*(x) \quad \text{where} \quad \hat{p}^*(x) := \prod_{m=1}^M \hat{p}_m(x) \quad (\text{H2-a})$$

and for each  $m \in \{1, \dots, M\}$   $\hat{p}_m(x)$  is the kernel density estimator of the probability density  $p_m(x)$  that has the form

$$\hat{p}_m(x) = \frac{1}{N_m h_m} \sum_{i=1}^{N_m} K\left(\frac{x - X_i^m}{h_m}\right). \quad (\text{H2-b})$$

Here  $X_1^m, X_2^m, \dots, X_{N_m}^m \sim p_m(x)$  are independent identically distributed random variables,  $K$  is a kernel density function, and  $h_m > 0$  is a bandwidth parameter.

The mean integrated square error of the estimator  $\hat{p}^*$  of the product  $p^*$  is defined by

$$\text{MISE}_{[\mathbf{N}, \mathbf{h}]} := \text{MISE}(p^*, \hat{p}^*(x)) = \mathbb{E} \int_{\mathbb{R}} (\hat{p}^*(x) - p^*(x))^2 dx \quad (2.2)$$

where we use the notation  $\mathbf{h} = (h_m)_{m=1}^M$  and  $\mathbf{N} = (N_m)_{m=1}^M$ . We also use the following convention for the bias and variance of estimators  $\hat{p}_m$

$$\begin{aligned} \text{bias}(\hat{p}^*(x)) &= \mathbb{E}[\hat{p}^*(x)] - p^*(x) \\ \text{bias}(\hat{p}_m(x)) &= \mathbb{E}[\hat{p}_m(x)] - p_m(x), \quad m \in \{1, \dots, M\}. \end{aligned} \quad (2.3)$$

We assume that the kernel density function  $K$  and probability densities functions  $p_1, \dots, p_M$  satisfy the following hypotheses:

**(H3)**  $K$  is positive, bounded, normalized, and its first moment is zero, that is

$$0 \leq K(t) \leq C, \quad \int_{\mathbb{R}} K(t) dt = 1, \quad \int_{\mathbb{R}} t K(t) dt = 0, \quad \int_{\mathbb{R}} K^2(t) dt < \infty \quad (2.4)$$

**(H4)** For each  $s \in \{0, 1, 2, 3\}$

$$k_s = \int_{\mathbb{R}} |t|^s K(t) dt < \infty. \quad (2.5)$$

**(H5)** For each  $m \in \{1, \dots, M\}$ ,  $s \in \{0, 1, 2, 3\}$  and density  $p_m \in C^3(\mathbb{R})$  there exists  $C_{m,s} \geq 0$  such that

$$|p_m^{(s)}(x)| < C_{m,s} \quad \text{for all } x \in \mathbb{R}. \quad (2.6)$$

**(H6)** For each  $m \in \{1, \dots, M\}$  and  $s \in \{0, 1, 2, 3\}$

$$\int_{\mathbb{R}} |p_m^{(s)}(x)| dx = I_{m,s} < \infty. \quad (2.7)$$

(H7) Let  $U = \{x : p^*(x) > 0\}$ . For each  $\tau_1, \tau_2, \dots, \tau_M > 0$  with  $\sum_{i=1}^M \tau_i = 1$  there exists  $x_0 \in U$  such that

$$\frac{d^2}{dx^2} \left( p_1^{\tau_1}(x_0) p_2^{\tau_2}(x_0) \cdots p_M^{\tau_M}(x_0) \right) > 0. \quad (2.8)$$

(H8) Functions

$$\mathbf{N}(n) = \{N_1(n), N_2(n), N_3(n), \dots, N_M(n)\} : \mathbb{N} \rightarrow \mathbb{N}^M$$

$$\mathbf{h}(n) = \{h_1(n), h_2(n), \dots, h_M(n)\} : \mathbb{N} \rightarrow \mathbb{R}_{++}^M$$

satisfy for all  $i \in \{1, 2, \dots, M\}$

$$\begin{aligned} D_1 &\leq \frac{N_i}{n} \leq D_2 \\ A_1 N_i(n)^{-\alpha_0} &\leq h_i(n) \leq A_2 N_i(n)^{-\alpha_0} \quad \text{for some } \alpha_0 \in (0, 1) \\ \lim_{n \rightarrow \infty} h_i(n) N_i(n) &= \infty. \end{aligned} \quad (2.9)$$

We also define  $\underline{N}(n) = \min_i N_i(n)$  and note that  $C_1 \|\mathbf{N}\| \leq \underline{N}(n) \leq C_2 \|\mathbf{N}(n)\|$ .

**Remark 2.1.** (H7) is a technical hypothesis which we employ in the proof of the uniqueness of the optimal bandwidth vector. We note that the hypothesis (H7) always holds for  $p^* \in C^3(\mathbb{R})$  with  $\{x : p^*(x) > 0\} = \mathbb{R}$ . In the general case when  $U \neq \mathbb{R}$  the hypothesis (H7) holds for instance whenever  $p^* \in C^\infty(\mathbb{R})$  or if each  $p_m$  decays fast enough near each boundary point of  $U$ ; for instance, if  $a$  is a boundary point of  $U$  and  $p_m$  behaves as  $e^{-\frac{1}{(x-a)^2}}$  near  $a$ .

### 3 Asymptotic analysis for MISE of the likelihood $\hat{p}^*(\mathbf{x}|\mathbf{y})$

We start with the observation that MISE can be expressed via the combination of bias and variance

$$\begin{aligned} \text{MISE}(p^*, \hat{p}^*) &= \mathbb{E} \int_{\mathbb{R}} (\hat{p}^*(x) - p^*(x))^2 dx \\ &= \int_{\mathbb{R}} \left( \text{bias}(\hat{p}^*(x), p^*(x)) \right)^2 dx + \int_{\mathbb{R}} \mathbb{V}(\hat{p}^*(x)) dx. \end{aligned} \quad (3.1)$$

In what follows we do the analysis of the bias, then that of variance and conclude with the section where we derive the formula for the optimal bandwidth vector.

#### 3.1 Bias expansion

Using the fact that  $\hat{p}_i(x)$ ,  $i = 1, \dots, M$  are independent, we obtain

$$\begin{aligned} \text{bias}(x) &= \mathbb{E}[\hat{p}^*(x)] - p^*(x) \\ &= \prod_{m=1}^M \mathbb{E}[\hat{p}_m](x) - \prod_{m=1}^M p_m(x) \\ &= \prod_{m=1}^M (\text{bias}_m(x) + p_m(x)) - \prod_{m=1}^M p_m(x) \end{aligned} \quad (3.2)$$

To simplify notation in (3.2) we shall employ the multiindex notation. Let  $\alpha$  be the multiindex with

$$\alpha = (\alpha_1, \alpha_2, \dots, \alpha_M) \quad \alpha_m \in \{0, 1\}.$$

Then the above formula can be rewritten as follows

$$\begin{aligned}
\text{bias}(x) &= \sum_{1 \leq |\alpha| \leq M} \prod_{m=1}^M \text{bias}_m^{\alpha_m}(x) (p_m(x))^{(1-\alpha_m)} \\
&= \sum_{m=1}^M \left[ \text{bias}_m(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] \\
&\quad + \sum_{2 \leq |\alpha| \leq M} \prod_{m=1}^M (\text{bias}_m(x))^{\alpha_m} (p_m(x))^{(1-\alpha_m)}.
\end{aligned} \tag{3.3}$$

Employing the above formula we prove the following lemma

**Lemma 3.1.** *Suppose hypotheses (H3)-(H6) hold. Then*

(i) *The bias can be expressed as*

$$\text{bias}(x) = \frac{k_2}{2} \sum_{m=1}^M \left[ h_m^2 p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] + E_b(x; \mathbf{h}) \tag{3.4}$$

where the error term  $E_b(x; \mathbf{h})$  satisfies the bounds

$$\begin{aligned}
|E_b(x; \mathbf{h})| &\leq E_\infty \|\mathbf{h}\|^3, \quad \forall x \in \mathbb{R} \\
\int_{\mathbb{R}} |E_b(x; \mathbf{h})| dx &\leq E_1 \|\mathbf{h}\|^3 \\
\int_{\mathbb{R}} |E_b(x; \mathbf{h})|^2 dx &\leq E_2 \|\mathbf{h}\|^6
\end{aligned} \tag{3.5}$$

(ii) *The square-integrated bias satisfies*

$$\int_{\mathbb{R}} \text{bias}^2(x) dx = \frac{k_2^2}{4} \int_{\mathbb{R}} \left[ \sum_{m=1}^M h_m^2 p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right]^2 dx + \mathcal{E}_b(\mathbf{h}) < \infty \tag{3.6}$$

with the error term satisfying

$$|\mathcal{E}_b(\mathbf{h})| \leq C_b \|\mathbf{h}\|^5 \tag{3.7}$$

where the constant  $C_b$  is independent of  $\mathbf{N}$  and  $\mathbf{h} \in \mathbb{R}_+^M$ .

*Proof.* According to (3.3) and (5.2) we have

$$\begin{aligned}
\text{bias}(x) &= \\
&= \frac{k_2}{2} \sum_{m=1}^M \left[ h_m^2 p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] + \sum_{m=1}^M \left[ E_{b,m} \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] \\
&\quad + \sum_{2 \leq |\alpha| \leq M} \prod_{m=1}^M \left( \frac{h_m^2 k_2}{2} p_m''(x) + E_{b,m} \right)^{\alpha_m} (p_m(x))^{(1-\alpha_m)}
\end{aligned}$$

We are computing bounds for

$$E_b(x; \mathbf{h}) = \sum_{m=1}^M \left[ E_{b,m} \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] + \sum_{2 \leq |\alpha| \leq M} \prod_{m=1}^M \left( \frac{h_m^2 k_2}{2} p_m''(x) + E_{b,m} \right)^{\alpha_m} (p_m(x))^{(1-\alpha_m)} \quad (3.8)$$

To simplify the derivations we separate the terms in (3.8) into two groups: terms with at least one multiple of  $E_{b,m}$  and terms free of  $E_{b,m}$ . We define the sets

$$A_m = \left\{ \alpha = (\alpha_j)_{j=1}^M : \alpha_m = 0 \text{ and } 1 \leq |\alpha| \leq (M-1) \right\} \quad (3.9)$$

and functions

$$P_m(x) = \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) + \sum_{\alpha \in A_m} \left[ \prod_{\substack{j=1 \\ j \neq m}}^M \left( \frac{h_j^2 k_2}{2} p_j''(x) + \mathbb{1}_{\{j>m\}} E_{b,j} \right)^{\alpha_j} (p_j(x))^{(1-\alpha_j)} \right]. \quad (3.10)$$

Here  $\mathbb{1}$  is the characteristic function. Consequently, the error term can be written as follows

$$E_b(x; \mathbf{h}) = \sum_{m=1}^M [E_{b,m} P_m(x)] + \sum_{2 \leq |\alpha| \leq M} \prod_{m=1}^M \left( \frac{h_m^2 k_2}{2} p_m''(x) \right)^{\alpha_m} (p_m(x))^{(1-\alpha_m)}. \quad (3.11)$$

Assuming that  $\|\mathbf{h}\|$  is bounded, (H5) and (5.2), we can conclude that there is a constant  $C_P$  so that

$$|P_m(x)| \leq C_P \text{ for any } x \in \mathbb{R} \text{ and } 1 \leq m \leq M$$

The first sum then can be easily bounded using (H5) and (5.2)

$$\sum_{m=1}^M |E_{b,m} P_m(x)| \leq \sum_{m=1}^M \left( \frac{C_{m,3} k_3 h_m^3}{6} C_P \right) \leq M \frac{\|\mathbf{h}\|^3 k_3}{6} C_3 C_P,$$

where  $C_j = \max_{1 \leq m \leq M} C_{m,j}$ .

The next sum in (3.11) contains terms that can be bounded as follows:

$$\left| \frac{h_m^2 k_2}{2} p_m''(x) \right| \leq \frac{\|\mathbf{h}\|^2 C_2 k_2}{2} \quad \text{and} \quad |p_m(x)| \leq C_0$$

Each one of these products will have at least two terms with  $p_m''(x)$  for some  $m$ , therefore

$$\left| \sum_{2 \leq |\alpha| \leq M} \prod_{m=1}^M \left( \frac{h_m^2 k_2}{2} p_m''(x) \right)^{\alpha_m} (p_m(x))^{(1-\alpha_m)} \right| \leq \frac{\|\mathbf{h}\|^4 C_2^2 k_2^2}{4} C_Q \quad (3.12)$$

for some appropriate constant  $C_Q$ . This implies the first inequality in (3.5):

$$|E_b(x; \mathbf{h})| \leq M \frac{\|\mathbf{h}\|^3 k_3}{6} C_3 C_P + \frac{\|\mathbf{h}\|^4 C_2^2 k_2^2}{4} C_Q \quad (3.13)$$

To prove  $L_1$  integrability we use conditions (H5), (H6), the expansion (3.11) and the second formula in (5.4)

$$\int_{\mathbb{R}} |E_b(x; \mathbf{h})| dx \leq M \frac{I_3 k_3 \|\mathbf{h}\|^3}{6} C_P + \frac{\|\mathbf{h}\|^2 k_2 I_2}{2} \cdot \frac{\|\mathbf{h}\|^2 k_2 C_2}{2} C_Q, \quad (3.14)$$

which proves the second estimate in (3.5).

Using the estimates obtained above, we conclude

$$\int_{\mathbb{R}} |E_b(x; \mathbf{h})|^2 dx \leq \sup_{\mathbb{R}} |E_b(x; \mathbf{h})| \cdot \int_{\mathbb{R}} |E_b(x; \mathbf{h})| dx \leq E_{\infty} \cdot E_1 \|\mathbf{h}\|^6$$

Finally,

$$\begin{aligned} \text{bias}^2(x) &= \frac{k_2^2}{4} \left[ \sum_{m=1}^M h_m^2 p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right]^2 + \\ &\quad + E_b(x; \mathbf{h}) k_2 \left[ \sum_{m=1}^M h_m^2 p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] + E_b^2(x; \mathbf{h}) \end{aligned}$$

Then Cauchy-Schwartz inequality leads to (3.6)  $\square$

### 3.2 Variance expansion

We next obtain an asymptotic formula for the variance of  $\hat{p}^*$ . For the proof of the lemma, we perform the following preliminary calculation

$$\begin{aligned} \mathbb{V}(\hat{p}^*(x)) &= \mathbb{E}[(\hat{p}^*(x))^2] - \left( \mathbb{E}[\hat{p}^*(x)] \right)^2 = \prod_{m=1}^M \mathbb{E}[\hat{p}_m^2] - \prod_{m=1}^M \mathbb{E}^2[\hat{p}_m] \\ &= \prod_{m=1}^M \left( \mathbb{V}[\hat{p}_m] + (p_m + \text{bias}[\hat{p}_m])^2 \right) - \prod_{m=1}^M (p_m + \text{bias}[\hat{p}_m])^2 \quad (3.15) \\ &= \sum_{1 \leq |\alpha| \leq M} \prod_{m=1}^M (\mathbb{V}[\hat{p}_m])^{\alpha_m} (p_m + \text{bias}[\hat{p}_m])^{2(1-\alpha_m)} \end{aligned}$$

**Lemma 3.2.** *Let hypotheses (H3)-(H6) hold. Then*

(i) *The variation of  $\hat{p}^*$  is given by*

$$\mathbb{V}(\hat{p}^*(x)) = \left( \sum_{m=1}^M \left[ \frac{p_m}{N_m h_m} \prod_{\substack{k=1 \\ k \neq m}}^M p_k^2(x) \right] \right) \int_{\mathbb{R}} K^2(t) dt + E_V(x; \mathbf{N}, \mathbf{h}), \quad x \in \mathbb{R} \quad (3.16)$$

where the error term  $E_V(x; \mathbf{N}, \mathbf{h})$  satisfies the bounds

$$|\mathcal{E}_V(N, h)| := \left| \int_{\mathbb{R}} E_V(x) dx \right| = o\left(\frac{1}{N_{\min}}\right) \quad \text{where} \quad N_{\min} = \min_{m \in \{1, \dots, M\}} N_m. \quad (3.17)$$

*Proof.* According to (3.15) we have

$$\begin{aligned} \mathbb{V}(\hat{p}^*(x)) &= \\ &= \sum_{1 \leq |\alpha| \leq M} \prod_{m=1}^M \left( \frac{p_m(x)}{N_m h_m} \int_{\mathbb{R}} K^2(t) dt + E_{V,m} \right)^{\alpha_m} (p_m + \text{bias}[\hat{p}_m])^{2(1-\alpha_m)} \\ &= \sum_{1 \leq |\alpha| \leq M} \prod_{m=1}^M \left( \frac{p_m(x)}{N_m h_m} \int_{\mathbb{R}} K^2(t) dt + E_{V,m} \right)^{\alpha_m} (p_m^2 + 2p_m \text{bias}[\hat{p}_m] + \text{bias}^2[\hat{p}_m])^{(1-\alpha_m)} \quad (3.18) \end{aligned}$$



In a fashion similar to the previous proof, we separate the terms in (3.18). We single out the leading order terms, the terms with at least one multiple of  $E_{V,m}$ , the terms with multiples of  $\text{bias}[\widehat{p}_m]$  and the terms of the order  $o\left(\frac{1}{N_{min}h_{min}}\right)$ .

We define sets

$$\begin{aligned} A_m^0 &= \left\{ \alpha = (\alpha_j)_{j=1}^M : \alpha_m = 0 \text{ and } 0 \leq |\alpha| \leq (M-1) \right\} \\ B_m^1 &= \left\{ \alpha = (\alpha_j)_{j=1}^M : \alpha_m = 0 \text{ and } |\alpha| = 1 \right\} \end{aligned} \quad (3.19)$$

and functions

$$\begin{aligned} P_m^0(x) &= \sum_{\alpha \in A_m^0} \left[ \prod_{\substack{j=1 \\ j \neq m}}^M \left( \frac{p_m(x)}{N_m h_m} \int_{\mathbb{R}} K^2(t) dt + \mathbf{1}_{\{j>m\}} E_{V,m} \right)^{\alpha_m} (\mathbb{E}^2[\widehat{p}_m])^{(1-\alpha_m)} \right], \\ Q_m^1(x) &= \sum_{\alpha \in B_m^1} \left[ \prod_{\substack{j=1 \\ j \neq m}}^M \left( \frac{p_m(x)}{N_m h_m} \int_{\mathbb{R}} K^2(t) dt \right)^{\alpha_m} (\mathbb{E}^2[\widehat{p}_m])^{(1-\alpha_m)} \right], \end{aligned} \quad (3.20)$$

The variance expansion can be rewritten as

$$\begin{aligned} \mathbb{V}(\widehat{p}^*(x)) &= \\ &= \sum_{1 \leq |\alpha| \leq M} \prod_{m=1}^M \left( \frac{p_m(x)}{N_m h_m} \int_{\mathbb{R}} K^2(t) dt \right)^{\alpha_m} (p_m^2 + 2p_m \text{bias}[\widehat{p}_m] + \text{bias}^2[\widehat{p}_m])^{(1-\alpha_m)} \\ &\quad + \sum_{m=1}^M E_{V,m} P_m^0(x) \\ &= \sum_{m=1}^M \left( \frac{p_m(x)}{N_m h_m} \int_{\mathbb{R}} K^2(t) dt \right) \prod_{\substack{j=1 \\ j \neq m}}^M p_m^2(x) \\ &\quad + \sum_{m=1}^M \text{bias}[\widehat{p}_m] (2p_m(x) + \text{bias}[\widehat{p}_m]) Q_m^1(x) \\ &\quad + \sum_{2 \leq |\alpha| \leq M} \prod_{m=1}^M \left( \frac{p_m(x)}{N_m h_m} \int_{\mathbb{R}} K^2(t) dt \right)^{\alpha_m} (\mathbb{E}^2[\widehat{p}_m])^{(1-\alpha_m)} \\ &\quad + \sum_{m=1}^M E_{V,m} P_m^0(x) \end{aligned} \quad (3.21)$$

Based on definitions of functions  $P_m^0(x)$  and  $Q_m^1(x)$ , hypotheses (H5) and (H6) we have the following bounds

$$\begin{aligned} \mathbb{E}[\widehat{p}_m] &\leq C_{\mathbb{E}} \\ |P_m^0(x)| &\leq C_P \\ |Q_m^1(x)| &\leq C_Q \frac{1}{N_{min} h_{min}} \end{aligned}$$

Therefore

$$\begin{aligned}
& \int_{\mathbb{R}} |E_V(x)| dx \\
& \leq \sum_{m=1}^M \left( 2C_0 + \frac{\|\mathbf{h}\|^2 C_2 k_2}{2} \right) \frac{C_Q}{N_{\min} h_{\min}} \int_{\mathbb{R}} |\text{bias}[\widehat{p}_m]| dx \\
& \quad + \frac{1}{N_{\min}^2 h_{\min}^2} \sum_{2 \leq |\alpha| \leq M} \left( \frac{1}{N_{\min}^2 h_{\min}^2} \right)^{(|\alpha|-2)} C_{\mathbb{E}}^{(M-|\alpha|)} \\
& \quad + \frac{M \cdot C_P}{N_{\min}}
\end{aligned} \tag{3.22}$$

This leads directly to (3.17).  $\square$

### 3.3 AMISE formula and optimal bandwidth vector

With the lemmas above we can derive the main result

**Theorem 3.3.** *Let hypotheses (H3)-(H7) hold. Then for every  $\mathbf{N} \in \mathbb{N}^M$  and  $\mathbf{h} \in \mathbb{R}_+^M$*

$$\begin{aligned}
\text{MISE}(p^*, \widehat{p}^*) &= \frac{k_2^2}{4} \int_{\mathbb{R}} \left( \sum_{m=1}^M \left[ h_m^2 p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] \right)^2 dx + \\
& \quad + \int_{\mathbb{R}} \left( \sum_{m=1}^M \left[ \frac{p_m}{N_m h_m} \prod_{\substack{k=1 \\ k \neq m}}^M p_k^2 \right] \right) dx \int_{\mathbb{R}} K^2(t) dt + \\
& \quad + \mathcal{E}_{\text{MISE}}(N, \mathbf{h})
\end{aligned} \tag{3.23}$$

where  $\mathcal{E}_{\text{MISE}}(\mathbf{N}, \mathbf{h}) = \mathcal{E}_b(\mathbf{N}, \mathbf{h}) + \mathcal{E}_V(\mathbf{N}, \mathbf{h})$  satisfies

$$\mathcal{E}_{\text{MISE}}(N, \mathbf{h}) = o\left(\|\mathbf{h}\|^4 + \frac{1}{N_{\min} h_{\min}}\right) \tag{3.24}$$

as  $\mathbf{h} \rightarrow \infty$ ,  $\mathbf{N} \rightarrow \infty$ , and  $(\|\mathbf{N}\| \|\mathbf{h}\|)^{-1} \rightarrow \infty$ .

*Proof.* The result follows from Lemma 3.1, Lemma 3.2, and formula (3.1)  $\square$

We next prove an elementary lemma that we use later.

**Lemma 3.4.** *Let  $\{p_m\}_{m=1}^M$  satisfy hypotheses (H5)-(H7). Let*

$$\Gamma = \left\{ \tau = (\tau_1, \tau_2, \dots, \tau_M) : \tau_m > 0 \text{ for each } m \in 1, \dots, M \text{ and } \sum_{m=1}^M \tau_m = 1 \right\}$$

and let

$$\nu(\tau) := \int_{\mathbb{R}} \left( \sum_{m=1}^M \left[ \tau_m p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] \right)^2 dx, \quad \tau \in \Gamma. \tag{3.25}$$

Then

$$\inf_{\tau \in \Gamma} \nu(\tau) > 0. \tag{3.26}$$

*Proof.* Let us argue by contradiction. Suppose that  $\nu(\tau) = 0$  for some  $\tau \in \Gamma$ . Then for all  $x \in U = \{x : p^*(x) > 0\}$  we must have

$$\begin{aligned}
\frac{d^2}{dx^2} \prod_{m=1}^M p_m^{\tau_m}(x) &= \\
&= \sum_{k=1}^M \tau_k p_k'' p_k^{\tau_k-1} \prod_{\substack{m=1 \\ m \neq k}}^M p_m^{\tau_m} + \sum_{k=1}^M \tau_k (\tau_k - 1) (p_k')^2 p_k^{\tau_k-2} \prod_{\substack{m=1 \\ m \neq k}}^M p_m^{\tau_m} \\
&\quad + \sum_{k=1}^M \tau_k p_k' p_k^{\tau_k-1} \left( \sum_{\substack{j=1 \\ j \neq k}}^M \tau_j p_j' p_j^{\tau_j-1} \prod_{\substack{m=1 \\ m \neq k, j}}^M p_m^{\tau_m} \right) = \\
&= \left( \prod_{r=1}^M p_r^{\tau_r-1} \right) \sum_{k=1}^M \tau_k p_k'' \prod_{\substack{m=1 \\ m \neq k}}^M p_m \\
&\quad - \left( \prod_{r=1}^M p_r^{\tau_r-2} \right) \sum_{k=1}^M \sum_{\substack{j=1 \\ j \neq k}}^M \tau_k \tau_j \left[ (p_k')^2 \prod_{\substack{m=1 \\ m \neq k}}^M p_m^2 - p_k' p_k p_j' p_j \prod_{\substack{m=1 \\ m \neq k, j}}^M p_m^2 \right]
\end{aligned}$$

Since the integrand in the formula for  $\nu(\tau)$  is nonnegative,  $\nu(\tau) = 0$  implies that the first term in the last two lines vanishes. Thus, rearranging some terms, we obtain

$$\begin{aligned}
\frac{d^2}{dx^2} \prod_{m=1}^M p_m^{\tau_m}(x) &= \\
&= - \left( \prod_{r=1}^M p_r^{\tau_r-2} \right) \sum_{k=1}^{M-1} \sum_{j=k+1}^M \tau_k \tau_j \prod_{\substack{m=1 \\ m \neq k, j}}^M p_m^2 [(p_k')^2 p_j^2 - 2p_k' p_k p_j' p_j + p_k^2 (p_j')^2] \\
&= - \left( \prod_{r=1}^M p_r^{\tau_r-2} \right) \sum_{k=1}^{M-1} \sum_{j=k+1}^M \tau_k \tau_j [p_k' p_j - p_k p_j']^2 \prod_{\substack{m=1 \\ m \neq k, j}}^M p_m^2 \leq 0
\end{aligned}$$

for all  $x \in U$ . This, contradicts the hypotheses (H7) and finishes the proof.  $\square$

Theorem 3.3 allows us to state that the choice of bandwidth vector  $\mathbf{h}$  that minimizes  $\text{MISE}(p^*, \hat{p}^*)$  is asymptotically indistinguishable from the vector  $\mathbf{h}$  that minimizes the following functional

$$\begin{aligned}
\text{AMISE}[p^*, \hat{p}^*; \mathbf{N}, \mathbf{h}] &= \frac{k_2^2}{4} \int_{\mathbb{R}} \left( \sum_{m=1}^M \left[ h_m^2 p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] \right)^2 dx + \\
&\quad + \int_{\mathbb{R}} \left( \sum_{m=1}^M \left[ \frac{p_m}{N_m h_m} \prod_{\substack{k=1 \\ k \neq m}}^M p_k^2 \right] \right) dx \int_{\mathbb{R}} K^2(t) dt
\end{aligned} \tag{3.27}$$

**Theorem 3.5 (minimization).** *The map  $\mathbf{h} \rightarrow \text{AMISE}(p^*, \hat{p}^*, \mathbf{h})$  is uniformly convex on*

$$\mathbb{R}_+^M = \{\mathbf{h} = (h_i)_{i=1}^M \in \mathbb{R}^M : h_i > 0 \text{ for } i = 1, \dots, M\}, \tag{3.28}$$

*and, consequently, has a unique minimizer  $\mathbf{h}_{\text{AMISE}} = \arg\min_{\mathbf{h} \in \mathbb{R}_+^M} \text{AMISE}(\cdot, \cdot, \mathbf{h})$ .*

*Proof.* The second integral in (3.27) is clearly a strictly convex sum of convex functionals in  $\mathbf{h}$ . Therefore it is sufficient to only show convexity of the term

$$J(\mathbf{h}) := \int_{\mathbb{R}} \left( \sum_{m=1}^M \left[ h_m^2 p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] \right)^2 dx \quad (3.29)$$

We can rewrite  $J(\mathbf{h})$ , using the functional  $\nu(\tau)$  defined in (3.25)

$$J(\mathbf{h}) = \|\mathbf{h}\|^4 \cdot \nu(\tau), \quad \text{where } \tau = \left( \frac{h_k^2}{\|\mathbf{h}\|^2} \right)_{k=1}^M \quad (3.30)$$

The function  $\nu(\tau)$  can be written as

$$\nu(\tau) = \tau Q \tau^T$$

where  $Q$  is a matrix with entries

$$Q_{i,j} = \int_{\mathbb{R}} \left( p_i''(x) \prod_{\substack{k=1 \\ k \neq i}}^M p_k(x) \right) \cdot \left( p_j''(x) \prod_{\substack{k=1 \\ k \neq j}}^M p_k(x) \right) dx$$

Since the functional  $\nu(\tau)$  has only non-negative values (for all  $\tau \in \mathbb{R}^M$ ), the matrix  $Q$  must be symmetric positive semi-definite. This guarantees convexity of the functional  $J(\mathbf{h})$  and therefore the mapping  $\mathbf{h} \rightarrow \text{AMISE}(\mathbf{h})$  is strictly convex. This, in turn, implies the uniqueness of the minimizer on  $\mathbb{R}_+^M$  as long as it exists.

The form of the function (3.27) and the result (3.26) imply that there exist positive constants  $D_1$  and  $D_2$  so that

$$\text{AMISE}(p^*, \hat{p}^*, \mathbf{h}) \geq D_1 \cdot \|\mathbf{h}\|^4 + \frac{D_2}{N_{max} h_{min}}. \quad (3.31)$$

The above relation implies that  $\text{AMISE}(\mathbf{h}) \rightarrow \infty$ , whenever  $\mathbf{h} \rightarrow \partial \mathbb{R}_+^M$  or  $\mathbf{h} \rightarrow \infty$ . This implies that the minimizer exists. This finishes the proof.  $\square$

### 3.4 Optimal parameters for symmetric likelihood

#### 3.4.1 General symmetric case

Theorem 3.5 insures the existence and uniqueness of an asymptotically optimal choice of bandwidth vector  $\mathbf{h}$ . Another conclusion, that is immediately apparent, is that local asymptotic minimization of MISE for each subset posterior estimate  $\hat{p}_m(x)$  using formulas (1.2) and (1.3) may lead to globally suboptimal estimate  $\hat{p}^*(x)$ , which has a significantly different formula for MISE. The procedure for locating the vector  $\mathbf{h}$  that minimizes (3.27) must be adjusted accordingly.

Unfortunately, locating the minimizer  $\mathbf{h}_{\text{AMISE}}$  requires solving nonlinear equations. Moreover, both MISE and AMISE require prior knowledge of unknown densities  $p_m$  in order to compute the minimizer. To go around this issue one can employ cross-validation techniques investigated in [3, 5]. This is beyond the scope of this article, and we limit the discussion to analysis of simple special cases and consideration of possible shortcut methods.

As a quick illustration of the importance of our result we consider special case when each machine works with i.i.d. subsets of samples of equal size. In other words, assume that

- $p_1 = p_2 = \dots = p_M$ .
- $N_1 = N_2 = \dots = N_M$ , that is,  $\mathbf{N} = (n, n, \dots, n)$ , for some  $n \in \mathbb{N}$

Due to the symmetry to find the optimal bandwidth it is sufficient to look for the minimizers among the bandwidth vectors of the form  $\mathbf{h} = (h, h, \dots, h)$ . In this ‘radial’ case  $\text{AMISE}[p^*, \hat{p}^*]$  becomes

$$\begin{aligned} \text{AMISE}[p^*, \hat{p}^*; \mathbf{N}, \mathbf{h}] &= \frac{k_2^2}{4} \int_{\mathbb{R}} \left( M h^2 p_1''(x) (p_1(x))^{(M-1)} \right)^2 dx + \\ &+ \int_{\mathbb{R}} \left( \frac{M}{nh} (p_1(x))^{2M-1} \right) dx \int_{\mathbb{R}} K^2(t) dt \end{aligned} \quad (3.32)$$

This expression is minimized by the bandwidth value  $\mathbf{h}^{\text{opt}} = (1, 1, \dots, 1) h_*^{\text{opt}}$  where

$$h_*^{\text{opt}} = (Mn)^{-1/5} k_2^{-2/5} \left( \frac{\int_{\mathbb{R}} (p_1(x))^{(2M-1)} dx \int_{\mathbb{R}} K^2(t) dt}{\int_{\mathbb{R}} p_1^{(2M-2)}(x) (p_1''(x))^2 dx} \right)^{1/5}. \quad (3.33)$$

The presence of extra factors in this formula shifts the minimizer to a location significantly different from (1.3).

### 3.4.2 Normal subset posterior densities

We next assume that all subsets of samples of  $x$  satisfy

- $p_m = \mathcal{N}(x, \mu, \sigma)$  is a normal distribution with the same mean and standard deviation for each  $m = 1, \dots, M$
- $N_1 = N_2 = \dots = N_M$ , that is,  $\mathbf{N} = (n, n, \dots, n)$ , for some  $n \in \mathbb{N}$ .

It is easy to show, using symmetry argument, that under our assumptions all components  $h_m$  of the vector  $\mathbf{h}_{\text{AMISE}}$  are equal (i.e.  $h_m = h$ ). Thus we perform the derivations taking advantage of this fact. In this simplified case we can derive the exact formula for  $\mathbf{h}_{\text{AMISE}}$ . This can be used as a replacement for the “rule of thumb” approximation commonly used in kernel density approximation implementations (see Silverman [25]).

Under our assumptions, we need to find a minimizer to

$$\begin{aligned} \text{AMISE}[p^*, \hat{p}^*; \mathbf{N}, \mathbf{h}] &= \frac{k_2^2}{4} \int_{\mathbb{R}} \left( M h^2 \frac{(x - \mu)^2 - \sigma^2}{\sigma^4} (\mathcal{N}(x, \mu, \sigma))^M \right)^2 dx + \\ &+ \int_{\mathbb{R}} \left( \frac{M}{nh} (\mathcal{N}(x, \mu, \sigma))^{2M} \right) dx \int_{\mathbb{R}} K^2(t) dt \end{aligned} \quad (3.34)$$

The minimizer of the (3.34) then is given by

$$\mathbf{h}_*^{\text{opt}} = (1, 1, \dots) h_*^{\text{opt}} \quad \text{with} \quad h_*^{\text{opt}} = \frac{2^{2/5} M^{3/10}}{(2M - 1)^{1/10} (4M^2 - 4M + 3)^{1/5}} \sigma n^{-1/5}. \quad (3.35)$$

Recall that  $n$  is the number of samples that each subset contains. Thus the total number of samples for all subsets is given by  $\|\mathbf{N}\|_1 = n \cdot M$ . In that case, letting  $M \rightarrow \infty$  we obtain

$$\lim_{M \rightarrow \infty} \left( \frac{\sigma}{2^{1/10} (nM)^{1/5}} \right)^{-1} \left( \frac{2^{2/5} M^{3/10}}{(2M - 1)^{1/10} (4M^2 - 4M + 3)^{1/5}} \sigma \cdot n^{-1/5} \right) = 1$$

and therefore

$$h_*^{\text{opt}} = (2^{-1/10} + O(M^{-1})) (nM)^{-1/5} \sigma \quad \text{as} \quad M \rightarrow \infty. \quad (3.36)$$

Setting  $M = 1$  in (3.35) yields the bandwidth vector

$$\mathbf{h}_0^{\text{opt}} = (1, 1, \dots) h_{M=1}^{\text{opt}} \quad \text{with} \quad h_{M=1} = \left( \frac{4}{3} \right)^{1/5} \sigma n^{-1/5} \quad (3.37)$$

where each component  $h_{M=1}^{\text{opt}}$  is the optimal bandwidth parameter for the individual subset posterior density estimator, a formula known as a “rule of thumb”. From the above analysis it follows that the choice of the bandwidth vector as  $\mathbf{h}_0^{\text{opt}}$ , as counterintuitive as it is, leads to a suboptimal approximation of  $\hat{p}^*(x)$ .

## 4 Asymptotic analysis of MISE for probability density

### 4.1 Re-normalization constant

In this section we consider the error that arises when one takes into account the re-normalization constant. Recall that by assumption

$$p(x) \propto p^*(x) \quad \text{where} \quad p^*(x) := \prod_{m=1}^M p_m(x)$$

where  $p_m(x)$ ,  $m \in \{1, \dots, M\}$  is a probability density function. Then we define

$$\lambda := \int p^*(x) dx > 0 \quad \text{and} \quad c := \lambda^{-1} \tag{4.1}$$

and obtain  $p(x) = cp^*(x)$ . For the estimator

$$\hat{p}(x) \propto \hat{p}^*(x) \quad \text{with} \quad \hat{p}^*(x) := \prod_{m=1}^M \hat{p}_m(x)$$

we similarly define

$$\hat{\lambda} := \int \hat{p}^*(x) dx > 0 \quad \text{and} \quad \hat{c} := \hat{\lambda}^{-1} \tag{4.2}$$

and hence  $\hat{p}(x) = \hat{c}\hat{p}^*(x)$ .

We are interested in the optimal bandwidth vector  $\mathbf{h} = (h)_{m=1}^M$  that optimizes the leading term of the mean integrated square error

$$\text{MISE}(\hat{p}, p) = \text{MISE}(\hat{c}\hat{p}^*, cp^*) = \mathbb{E} \int_{\mathbb{R}} (cp^*(x) - \hat{c}\hat{p}^*(x))^2 dx. \tag{4.3}$$

Observe that  $\hat{c}$  and  $\hat{p}^*$  are not independent and the previously performed analysis is not directly applicable. Moreover, we observe that the estimator of the renormalizing constant

$$\hat{c} = \left( \int \prod_{i=1}^M \hat{p}_i(x) dx \right)^{-1} < \infty$$

may in general have an infinite expectation. This may happen because the estimators in the above product may decay too quickly in  $x$  variable and the sets of  $x$  with the most ‘mass’ for each  $p_i$  may have no common intersection. This potentially may lead to small values of  $\hat{\lambda}$  and hence large  $\hat{c}$ . To avoid this situation one would need to chose the kernel  $K$  in appropriate way and establish the finiteness of the expectation of  $\hat{c}$ .

In this article we do not investigate this. Instead, we will show that one can replace MISE by an equivalent functional which is well-defined and finite on the whole sample space  $\Omega$  and that there exists a sequence of smaller sample subspaces  $\Omega_n$  with  $\mathbb{P}(\Omega_n) \rightarrow 1$ , on which the new functional is asymptotically *equivalent* to  $\text{MISE}(p, \hat{p})$  restricted to  $\Omega_n$ . We then analyze the new functional and investigate its optimal parameters.

## 4.2 Preliminary estimates

**Lemma 4.1 (covariance).** *Let  $\widehat{p}^*(x)$  be an estimator of the form (H2-a) where the vector of sample sizes  $\mathbf{N}(n)$  and bandwidth vector  $\mathbf{h}(n)$  satisfy (H8). Then*

$$\text{Cov}[\widehat{p}^*(x), \widehat{p}^*(y)] = \mathbb{E}[\widehat{p}^*(x)\widehat{p}^*(y)] - \mathbb{E}[\widehat{p}^*(x)]\mathbb{E}[\widehat{p}^*(y)] \quad (4.4)$$

satisfies the estimates

$$\begin{aligned} |\text{Cov}[\widehat{p}^*(x), \widehat{p}^*(y)]| &\leq \frac{C_{abs}}{\underline{N}(n)\underline{h}(n)} \leq \frac{\mu}{\|\mathbf{N}\|\|\mathbf{h}\|} \\ \left| \iint \text{Cov}[\widehat{p}^*(x), \widehat{p}^*(y)] dx dy \right| &\leq \frac{C_{int}}{\underline{N}(n)} \leq \frac{\mu}{\|\mathbf{N}\|} \end{aligned} \quad (4.5)$$

for some constants  $C_{abs}, C_{int}, \mu > 0$  independent of  $n$ .

*Proof.* We can expand the product as follows

$$\begin{aligned} \prod_{i=1}^M \mathbb{E}[\widehat{p}_i(x)\widehat{p}_i(y)] - \prod_{i=1}^M \mathbb{E}[\widehat{p}_i(x)]\mathbb{E}[\widehat{p}_i(y)] \\ = \sum_{j=1}^M \left( \mathbb{E}[\widehat{p}_j(x)\widehat{p}_j(y)] - \mathbb{E}[\widehat{p}_j(x)]\mathbb{E}[\widehat{p}_j(y)] \right) \left( \prod_{i=1}^{j-1} \mathbb{E}[\widehat{p}_i(x)\widehat{p}_i(y)] \right) \left( \prod_{i=j+1}^M \mathbb{E}[\widehat{p}_i(x)]\mathbb{E}[\widehat{p}_i(y)] \right) \end{aligned}$$

where the products with the top index smaller than the bottom index should be taken as having the value one.

We next observe that, according to (5.2), for each  $i \in \{1, \dots, M\}$

$$|\mathbb{E}[\widehat{p}_i(x)]\mathbb{E}[\widehat{p}_i(y)]| \leq \left( C_0 + \frac{C_2 k_2 h^2}{2} + \frac{C_3 k_3 h^3}{6} \right)^2.$$

Also Lemma 5.4 implies that

$$|\mathbb{E}[\widehat{p}_i(x)\widehat{p}_i(y)]| \leq \left( C_0 + \frac{C_2 k_2 h^2}{2} + \frac{C_3 k_3 h^3}{6} \right)^2 + \frac{C C_0}{nh} + \frac{1}{n} \left( C_1 C k_1 + \left( C_0 + \frac{C_2 k_2 h^2}{2} + \frac{C_3 k_3 h^3}{6} \right)^2 \right)$$

Then we conclude that for some  $C_{\mathbb{E}} \geq 0$

$$|\mathbb{E}[\widehat{p}_i(x)\widehat{p}_i(y)]|, |\mathbb{E}[\widehat{p}_i(x)]\mathbb{E}[\widehat{p}_i(y)]| \leq C_{\mathbb{E}} < \infty, \quad \text{for all } x, y \in \mathbb{R}.$$

Therefore, by Lemma 5.4 we obtain the estimate

$$|\text{Cov}[\widehat{p}^*(x), \widehat{p}^*(y)]| \leq M \left( \frac{C C_0}{nh} + \frac{1}{n} \left( C_1 C k_1 + \left( C_0 + \frac{C_2 k_2 h^2}{2} + \frac{C_3 k_3 h^3}{6} \right)^2 \right) \right) \cdot C_{\mathbb{E}}^{M-1}$$

which gives 4.5<sub>1</sub>.

The integral of  $\text{Cov}[\widehat{p}^*(x), \widehat{p}^*(y)]$  is also finite. Using the result of Lemma 5.4

$$\begin{aligned} \iint |\text{Cov}[\widehat{p}^*(x), \widehat{p}^*(y)]| dx dy \\ \leq C_{\mathbb{E}}^{M-1} \sum_{i=1}^M \iint |\mathbb{E}[\widehat{p}_j(x)\widehat{p}_j(y)] - \mathbb{E}[\widehat{p}_j(x)]\mathbb{E}[\widehat{p}_j(y)]| dx dy \\ \leq C_{\mathbb{E}}^{M-1} \sum_{i=1}^M \iint \left( \frac{1}{N_i} p_i(x) \frac{1}{h_i} K_2 \left( \frac{x-y}{h_i} \right) + |E_{\Pi, i}(x, y)| \right) dx dy \\ \leq C_{\mathbb{E}}^{M-1} \frac{M}{\underline{N}} \left( 2 + k_1 I_1 + \frac{I_2 k_2 \|\mathbf{h}\|^2}{2} + \frac{I_3 k_3 \|\mathbf{h}\|^3}{6} \right), \end{aligned}$$

Where at the last step we used the facts that  $\frac{1}{h}K_2\left(\frac{x-y}{h}\right)$  is a probability density function in  $y$  for any fixed  $x$  and  $p_i(x)$  is also a probability density function.  $\square$

**Lemma 4.2.** *Let  $\widehat{p}^*(x)$  be an estimator of the form (H2-a) where the vector of sample sizes  $\mathbf{N}(n)$  and bandwidth vector  $\mathbf{h}(n)$  satisfy (H8). Then following identity and the estimate holds*

$$\mathbb{V}(\widehat{\lambda} - \lambda) = \mathbb{V}\left(\int \widehat{p}^*(x) dx - \int p^*(x) dx\right) \leq \frac{C_{int}}{\underline{N}} \leq \frac{\mu}{\|\mathbf{N}\|} < \infty,$$

where  $C_{int}, \mu > 0$  are defined in (4.5).

*Proof.* Since  $\lambda$  is constant we have

$$\begin{aligned} \mathbb{V}(\widehat{\lambda} - \lambda) &= \mathbb{E}(\widehat{\lambda} - \mathbb{E}[\widehat{\lambda}])^2 \\ &= \mathbb{E}\left(\int_{\mathbb{R}} \widehat{p}^*(x) - \mathbb{E}[\widehat{p}^*(x)] dx\right)^2 \\ &= \mathbb{E}\left[\int (\widehat{p}^*(x) - \mathbb{E}[\widehat{p}^*(x)]) dx \cdot \int (\widehat{p}^*(y) - \mathbb{E}[\widehat{p}^*(y)]) dy\right] \\ &= \iint \left(\mathbb{E}[\widehat{p}^*(x)\widehat{p}^*(y)] - \mathbb{E}[\widehat{p}^*(x)]\mathbb{E}[\widehat{p}^*(y)]\right) dx dy \leq \frac{C_{int}}{\underline{N}}, \end{aligned}$$

where the last inequality is from Lemma 4.1.  $\square$

**Lemma 4.3.** *Let  $\widehat{p}^*(x)$  be an estimator of the form (H2-a) where the vector of sample sizes  $\mathbf{N}(n)$  and bandwidth vector  $\mathbf{h}(n)$  satisfy (H8). Then for any  $\alpha \in (0, 1]$*

$$\mathbb{P}\left(\left\{\omega : |\mathbb{E}\widehat{\lambda} - \widehat{\lambda}(\omega; \mathbf{N}(n), \mathbf{h}(n))| > \frac{\lambda}{\sqrt{2\underline{N}(n)^{\frac{1-\alpha}{2}}}}\right\}\right) \leq \frac{2C_{int}}{\lambda^2 \underline{N}(n)^\alpha}. \quad (4.6)$$

Moreover, for any  $\alpha$  satisfying

$$\max(0, 1 - 4\alpha_0) < \alpha < 1,$$

where  $\alpha_0$  is defined in (H8), we have

$$\mathbb{P}\left\{\left|\frac{\widehat{\lambda}}{\lambda} - 1\right| > \frac{1}{\underline{N}(n)^{\frac{1-\alpha}{2}}}\right\} \leq \frac{2C_{int}}{\lambda^2 \underline{N}(n)^\alpha}. \quad (4.7)$$

for all sufficiently large  $n$ .

*Proof.* By Lemma 4.2 and Chebychev inequality we obtain

$$\begin{aligned} &\mathbb{P}\left\{|\widehat{\lambda} - \mathbb{E}[\widehat{\lambda}]|^2 > \frac{\lambda^2}{2\underline{N}^{1-\alpha}}\right\} \\ &\leq \mathbb{P}\left\{|\widehat{\lambda} - \mathbb{E}[\widehat{\lambda}]|^2 > \mathbb{V}(\widehat{\lambda}) \frac{\lambda \underline{N}^\alpha}{2C_{int}}\right\} \leq \frac{2C_{int}}{\lambda^2 \underline{N}^\alpha}. \end{aligned}$$

Recall next that

$$|\mathbb{E}(\widehat{\lambda}) - \lambda| = \left|\int (\mathbb{E}[\widehat{p}^*(x)] - p^*(x)) dx\right| \leq \int |\text{bias}(\widehat{p}^*, p^*)| dx \leq C\|\mathbf{h}\|^2$$



where  $C$  is independent of  $\mathbf{h}$ . According to (H8) we have  $\|\mathbf{h}(n)\| \leq A\|\mathbf{N}\|^{-\alpha_0}$  for some  $\alpha_0 \in (0, 1)$ . Fix an arbitrary  $\alpha$  that satisfies

$$\max(0, 1 - 4\alpha_0) < \alpha < 1 \quad \text{so that} \quad 4\alpha_0 > 1 - \alpha.$$

Then

$$\|\mathbf{h}\|^2 \|\mathbf{N}\|^{\frac{1-\alpha}{2}} \leq A \|\mathbf{N}\|^{-2\alpha_0} \|\mathbf{N}\|^{\frac{1-\alpha}{2}} = \|\mathbf{N}\|^{-\frac{4\alpha_0 + (1-\alpha)}{2}} \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

Thus there exists  $n_0$  such that

$$C\|\mathbf{h}(n)\|^2 < \frac{\lambda}{4} \underline{N}^{-\frac{(1-\alpha)}{2}} \quad \text{for all } n > n_0. \quad (4.8)$$

By the triangle inequality we have

$$|\widehat{\lambda} - \mathbb{E}\widehat{\lambda}| > |\widehat{\lambda} - \lambda| - |\lambda - \mathbb{E}\widehat{\lambda}| > |\widehat{\lambda} - \lambda| - \frac{\lambda}{4} \underline{N}^{-\frac{(1-\alpha)}{2}}$$

and hence for every

$$\omega_0 \in \left\{ \omega : |\widehat{\lambda}(\omega) - \lambda| > \frac{\lambda}{\underline{N}^{\frac{1-\alpha}{2}}} \right\} \quad (4.9)$$

we have

$$|\widehat{\lambda}(\omega_0) - \mathbb{E}\widehat{\lambda}| > |\widehat{\lambda}(\omega_0) - \lambda| - \frac{\lambda}{4} \underline{N}^{-\frac{(1-\alpha)}{2}} > \frac{3\lambda}{4} \underline{N}^{-\frac{(1-\alpha)}{2}} > \frac{\lambda}{\sqrt{2}} \underline{N}^{-\frac{(1-\alpha)}{2}}. \quad (4.10)$$

Then (4.9) and (4.10) we obtain

$$\left\{ \omega : |\widehat{\lambda}(\omega) - \lambda| > \frac{\lambda}{\underline{N}^{\frac{1-\alpha}{2}}} \right\} \subset \left\{ \omega : |\widehat{\lambda}(\omega) - \mathbb{E}\widehat{\lambda}| > \frac{\lambda}{\sqrt{2} \underline{N}^{\frac{1-\alpha}{2}}} \right\}$$

and hence

$$\begin{aligned} & \mathbb{P} \left\{ \omega : |\widehat{\lambda}(\omega) - \lambda| > \frac{\lambda}{\underline{N}^{\frac{1-\alpha}{2}}} \right\} \\ & \leq \mathbb{P} \left\{ \omega : |\widehat{\lambda}(\omega) - \mathbb{E}\widehat{\lambda}| > \frac{\lambda}{\sqrt{2} \underline{N}^{\frac{1-\alpha}{2}}} \right\} \leq \frac{2C_{int}}{\lambda^2 \underline{N}^\alpha}. \end{aligned} \quad (4.11)$$

□

### 4.3 Functional equivalent to MISE

**Definition 4.4.** Let  $\mathbf{h}(n)$  and  $\mathbf{N}(n)$  satisfy (H9) and let  $\max(0, 1 - 4\alpha_0) < \alpha < 1$  be fixed. We set

$$\Omega_n = \left\{ \omega \in \Omega : |\widehat{\lambda} - \lambda| \leq \frac{\lambda}{\underline{N}^{\frac{1-\alpha}{2}}} \right\}$$

and notice that

$$\mathbb{P}(\Omega_n) \geq 1 - \frac{C}{\underline{N}^\alpha} \quad (4.12)$$

**Definition 4.5.** We define the distance functional by

$$\begin{aligned} \overline{\text{MISE}}[p, \widehat{p}] &= c^2 \left( \int \text{bias}[p^*, \widehat{p}^*] dx \right)^2 \int (cp^*(x))^2 dx \\ &+ c^2 \int \text{bias}^2[p^*, \widehat{p}^*] + \mathbb{V}[\widehat{p}^*] dx \\ &- 2c^2 \iint \text{bias}[\widehat{p}^*(y)] \text{bias}[\widehat{p}^*(x)] cp^*(x) dx dy \end{aligned} \quad (4.13)$$

We also define the restriction of  $\text{MISE}[p, \hat{p}]$  to the set  $\Omega_n$  as

$$\begin{aligned} \text{MISE}[p(x), \hat{p}(x, \omega) | \omega \in \Omega_n] &= \mathbb{E} \left[ \int (p(x) - \hat{p}(x, \omega))^2 dx \middle| \omega \in \Omega_n \right] \\ &= \int_{\Omega_n} \int (p(x) - \hat{p}(x, \omega))^2 dx d\mathbb{P}(\omega) \end{aligned} \quad (4.14)$$

**Proposition 4.6.** *The functional  $\overline{\text{MISE}}$  is asymptotically equivalent to  $\text{MISE}$  on spaces  $\Omega_n$  uniformly in  $n$ , that is*

$$\lim_{\|\mathbf{N}(n)\| \rightarrow \infty} \frac{\overline{\text{MISE}}[p(x), \hat{p}(x) | \Omega_n]}{\text{MISE}[p(x), \hat{p}(x, \omega) | \Omega_n]} = 1 \quad (4.15)$$

Remark: It must be noted that the equivalence is established on the smaller event space  $\Omega_n$ , which excludes the events for which  $\lambda$  values are too small. However, the functional  $\overline{\text{MISE}}$  is defined for the whole space  $\Omega$ . Our subsequent derivations are formulated for unrestricted  $\overline{\text{MISE}}$ .

The next logical step is to extract the leading order terms from the functional  $\overline{\text{MISE}}$ . In other words, we drop higher order terms from the quantities like  $\text{bias}[p^*, \hat{p}^*]$  and  $\mathbb{V}[\hat{p}^*]$ . Since the expansions have been previously derived we simply replace the terms and show that the dropped terms have higher order. We present these result in the lemma below.

**Lemma 4.7.** *The distance functional  $\overline{\text{MISE}}$  can be represented as*

$$\overline{\text{MISE}}[p(x), \hat{p}(x)] = \overline{\text{AMISE}}[p(x), \hat{p}(x)] + \mathcal{E}(\mathbf{N}, \mathbf{h}) \quad (4.16)$$

where

$$\begin{aligned} \overline{\text{AMISE}}[p(x), \hat{p}(x)] &:= \frac{c^2 k_2^2}{4} \left( \int \sum_{m=1}^M \left[ h_m^2 p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] dx \right)^2 \int (cp^*(x))^2 dx \\ &+ \frac{c^2 k_2^2}{4} \int \left( \sum_{m=1}^M \left[ h_m^2 p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] \right)^2 dx + \int_{\mathbb{R}} \left( \sum_{m=1}^M \left[ \frac{p_m}{N_m h_m} \prod_{\substack{k=1 \\ k \neq m}}^M p_k^2(x) \right] \right) dx \int_{\mathbb{R}} K^2(t) dt \\ &- \frac{c^2 k_2^2}{2} \iint \left( \sum_{m=1}^M \left[ h_m^2 p_m''(y) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(y) \right] \right) \left( \sum_{m=1}^M \left[ h_m^2 p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^M p_k(x) \right] \right) cp^*(x) dx dy \end{aligned} \quad (4.17)$$

and the error term  $\mathcal{E}(\mathbf{N}, \mathbf{h})$  satisfies

$$|\mathcal{E}(\mathbf{N}, \mathbf{h})| = o\left(\|\mathbf{h}\|^4 + \frac{1}{\|\mathbf{N}\| \|\mathbf{h}\|}\right) \quad (4.18)$$

as  $\mathbf{h} \rightarrow \infty$ ,  $\mathbf{N} \rightarrow \infty$ , and  $(\|\mathbf{N}\| \|\mathbf{h}\|)^{-1} \rightarrow \infty$ .

*Proof.* The result follows from Lemma 3.1, Lemma 3.2, and formula (3.1)  $\square$

#### 4.4 $\overline{\text{AMISE}}$ optimization for normal subset posterior densities

Let us assume that all subsets of samples of  $x$  satisfy

- $p_m = \mathcal{N}(x, \mu, \sigma)$  is a normal distribution with the same mean and standard deviation for each  $m = 1, \dots, M$

- $N_1 = N_2 = \dots = N_M$ , that is,  $\mathbf{N} = (n, n, \dots, n)$ , for some  $n \in \mathbb{N}$ .

Again, using symmetry argument, we look for the minimizer on the set of positive vectors  $\mathbf{h} = (h, h, \dots, h)$ . Under the above assumptions then we conclude that the minimizer of (4.17) is given by

$$\mathbf{h}^{\text{opt}} = (1, 1, \dots, 1)h^{\text{opt}} \quad \text{with} \quad h^{\text{opt}} = \frac{2^{2/5}M^{3/10}}{3^{1/5}(2M-1)^{1/10}}\sigma n^{-1/5}. \quad (4.19)$$

Recall that  $n$  is the number of samples that each subset contains and hence the total number of samples for all subsets is given by  $\|\mathbf{N}\|_1 = n \cdot M$ . Thus, letting  $M \rightarrow \infty$  we obtain

$$\lim_{M \rightarrow \infty} \left( \left( \frac{8}{9} \right)^{1/10} \sigma (nM)^{-1/5} \right)^{-1} \left( \frac{2^{2/5}M^{3/10}}{3^{1/5}(2M-1)^{1/10}} \sigma \cdot (nm)^{-1/5} \right) = 1$$

and therefore

$$h^{\text{opt}} = \left( (8/9)^{1/10} + O(M^{-1}) \right) (nM)^{-1/5} \sigma \quad \text{as} \quad M \rightarrow \infty. \quad (4.20)$$

Setting  $M = 1$  in (4.19) we once again obtain the bandwidth vector

$$\mathbf{h}_0^{\text{opt}} = (1, 1, \dots)h_{M=1}^{\text{opt}} \quad \text{with} \quad h_{M=1}^{\text{opt}} = \left( \frac{4}{3} \right)^{1/5} \sigma n^{-1/5}$$

where each component  $h_{M=1}^{\text{opt}}$  is the optimal bandwidth parameter for the individual subset posterior density estimator. Thus the ‘intuitive’ choice of the bandwidth vector as  $\mathbf{h}_0^{\text{opt}}$  leads to a suboptimal approximation of  $\hat{p}(x)$ .

**Remark 4.8.** The above analysis indicates that that for large number of partitions the optimal bandwidth parameter  $\mathbf{h}^{\text{opt}}$  for full set posterior density estimation differs from the case of likelihood estimation (3.36) by only a constant multiple.

#### 4.5 Numerical experiments with normal subset posterior densities

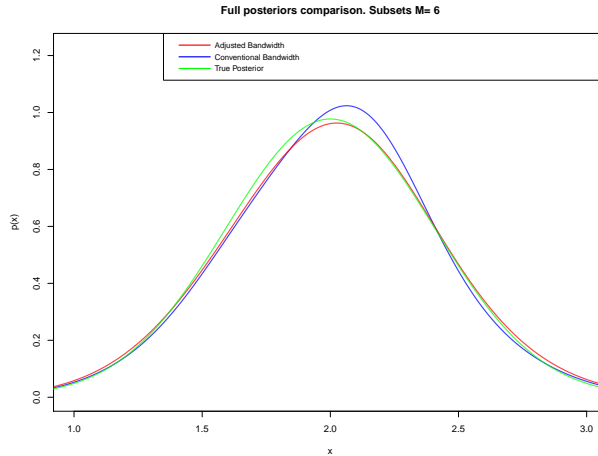


Figure 1: Graphs of  $p(x)$ -(red),  $\hat{p}(x; \mathbf{N}, \mathbf{h}_0^{\text{opt}})$ -(blue) and  $\hat{p}(x; \mathbf{N}, \mathbf{h}^{\text{opt}})$ -(green).

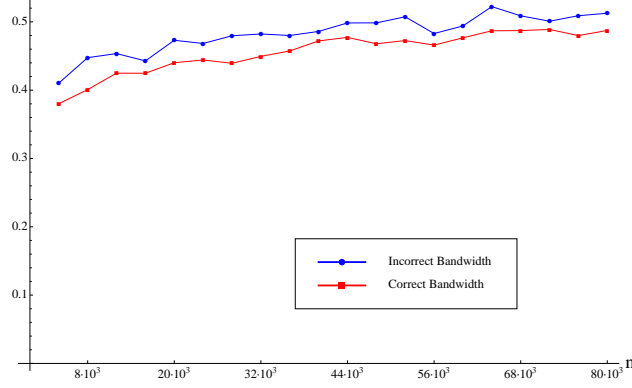


Figure 2:  $\text{MISE}(\mathbf{N}(n)) \times \|\mathbf{N}(n)\|_1^{4/5}$  for  $\mathbf{h}^{\text{opt}}$  and  $\mathbf{h}_0^{\text{opt}}$ .

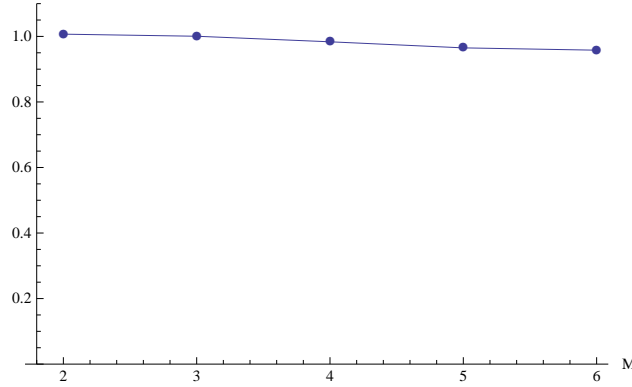


Figure 3: The ratio  $h^{\text{opt}}$  to numerically computed  $h_{\text{MISE}}^{\text{opt}}$ .

We conduct several numerical experiments to verify that the theoretical computations do predict the expected error accurately. We generate normally distributed data then produce two version of KDE estimators: one with bandwidth parameter  $h_m$  chosen to optimize error locally for the subset posterior, and another with bandwidth chosen according to formula derived in (4.19). Figure 1 contains graphs for typical posterior density functions generated with the numerical method of Conlon and Miroshnikov [16]. The red curve lies closer to the green curve, which results in a smaller error. Naturally, such comparisons are not legitimate, as both red and blue curve have an element of randomness in them. We can, though, compare the average errors for different densities by running the simulation multiple times and averaging the resulting errors (see Figure 2). Evidently, the error oscillates around a constant level, which verifies the conclusion that its order of decay is  $O(N^{-4/5})$ , also, the red curve that corresponds to the average error with adjusted bandwidth is consistently lower than the error curve (green) for the traditional choice of bandwidth. To verify that the formula (4.19) does indeed specify the optimal choice of bandwidth, we ran the simulation varying the bandwidth over a range of values. We then locate the values of  $h^{\text{opt}}$  for each combination of number of samples  $\mathbf{N}$  and the number of subset posterior subdivisions  $M$ , where the minimum of the (estimated) mean square error is achieved. The ratio of  $h^{\text{opt}}/h_{\text{MISE}}^{\text{opt}}$  is plotted on Figure 3 against the theoretically predicted values (squares).

## 5 Appendix

### 5.1 Kernel density estimators and asymptotic error analysis

In this section we will use the following notation. The function  $f$  denotes a probability density and its kernel density estimator is given by

$$\hat{f}(x; X_1, X_2, \dots, X_N, h) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x - X_i}{h}\right). \quad (5.1)$$

where  $X_1, X_2, \dots, X_n \sim f$  are i.i.d. samples.

**Lemma 5.1 (bias expansion).** *Let  $K$  satisfy (H3) and (H4). Let  $f$  be a probability density function satisfying (H5) and (H6). Let  $\hat{f}_{n,h}(x)$  be an estimation of  $f$  given by (5.1). Then*

(i) *bias( $\hat{f}_{n,h}$ ) is given by*

$$\begin{aligned} [\text{bias}(\hat{f}_{n,h})](x) &= \\ &= \mathbb{E}[\hat{f}_{n,h}(x)] - f(x) = \frac{h^2 k_2 f''(x)}{2} + [E_b(f, K)](x; h) \end{aligned} \quad (5.2)$$

where

$$E_b(x; h) := \int_{\mathbb{R}} K(t) \left( \int_x^{x-ht} \frac{f'''(z)(x - ht - z)^2}{2} dz \right) dt. \quad (5.3)$$

(ii) *For all  $n \geq 1$  and  $h > 0$  the term  $E_b(\cdot; n, h)$  satisfies the bounds*

$$\begin{aligned} |E_b(x; h)| &\leq \frac{C_3 k_3}{6} h^3, \quad x \in \mathbb{R} \\ \int_{\mathbb{R}} |E_b(x; h)| dx &\leq I_3 \frac{k_3}{6} h^3 \\ \int_{\mathbb{R}} |E_b(x; n, h)|^2 dx &\leq I_3 \frac{C_3 k_3^2}{36} h^6 \end{aligned} \quad (5.4)$$

(iii) *The square-integrated bias( $\hat{f}_{n,k}$ ) satisfies*

$$\int_{\mathbb{R}} \text{bias}^2(\hat{f}_{n,k}) dx = \frac{h^4 k_2^2}{4} \int_{\mathbb{R}} (f''(x))^2 dx + \mathcal{E}_b(n, h) < \infty \quad (5.5)$$

with

$$|\mathcal{E}_b(n, h)| \leq \left( k_2 C_2 + \frac{C_3 k_3}{6} h \right) \frac{h^5}{6} I_3 k_3 \quad (5.6)$$

for all  $n \geq 1$  and  $h > 0$ .

*Proof.* Using (5.1) and the fact that  $X_i, i = 1, \dots, n$  are i.i.d. we obtain

$$\begin{aligned} \text{bias}_{n,h}(x) &= \mathbb{E}[\hat{f}_{n,h}(x)] - f(x) = \\ &= \frac{1}{h} \mathbb{E} \left[ K\left(\frac{x - X_1}{h}\right) \right] - f(x) \\ &= \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{x - y}{h}\right) f(y) dy - f(x) \\ &= \int_{\mathbb{R}} K(t) (f(x - ht) - f(x)) dt \end{aligned}$$

where we used the substitution  $t = (x - y)/h$ . Employing Taylor's Theorem with an error term in integral form and using (H3) we get

$$\begin{aligned} \text{bias}_{n,h}(x) &= \int_{\mathbb{R}} K(t) \left( -htf'(x) + \frac{h^2 t^2}{2} f''(x) + \int_x^{x-ht} \frac{f'''(z)(x-ht-z)^2}{2} dz \right) dt \\ &= \frac{h^2 f''(x)}{2} \int_{\mathbb{R}} t^2 K(t) dt + \int_{\mathbb{R}} K(t) \left( \int_x^{x-ht} \frac{f'''(z)(x-ht-z)^2}{2} dz \right) dt \end{aligned}$$

which proves (i).

By (H4) we have

$$|E_b(x; n, h)| \leq C_3 \left( \int_{\mathbb{R}} K(t) \left| \int_x^{x-ht} \frac{(x-ht-z)^2}{2} dz \right| dt \right) = \frac{C_3 k_3}{6} h^3 \quad (5.7)$$

and by (H6), using the substitution  $\alpha = x - ht - z$  and employing Tonelli's Theorem, we obtain

$$\begin{aligned} &\int_{\mathbb{R}} |E_b(x; n, h)| dx \\ &\leq \int_{\mathbb{R}} \int_{\mathbb{R}} K(t) \int_{x-\frac{h}{2}(|t|+t)}^{x+\frac{h}{2}(|t|-t)} \frac{|f'''(z)|(x-ht-z)^2}{2} dz dt dx \\ &= \int_{\mathbb{R}} K(t) \int_{-\frac{h}{2}(|t|+t)}^{\frac{h}{2}(|t|-t)} \left( \int_{\mathbb{R}} |f'''(x-ht-\alpha)| d\alpha \right) \frac{\alpha^2}{2} d\alpha dt \\ &\leq I_3 \int_{\mathbb{R}} K(t) \left( \int_{-\frac{h}{2}(|t|-t)}^{\frac{h}{2}(|t|+t)} \frac{\alpha^2}{2} d\alpha \right) dt = \frac{h^3}{6} I_3 k_3. \end{aligned} \quad (5.8)$$

Thus, combining the two bounds above we conclude

$$\int_{\mathbb{R}} |E_b(x; n, h)|^2 dx \leq \frac{C_3 k_3}{6} h^3 \int_{\mathbb{R}} |E_b(x; n, h)| dx \leq I_3 \frac{C_3 k_3^2}{36} h^6.$$

Observe that

$$\text{bias}^2(\widehat{f}_{n,h})(x) = \frac{h^4 k_2^2}{4} (f''(x))^2 + h^2 k_2 f''(x) E_b(x; n, h) + E_b^2(x; n, h). \quad (5.9)$$

By (H5), (5.7) and (5.8)

$$\begin{aligned} |\mathcal{E}_b(n, h)| &:= \left| \int_{\mathbb{R}} \left( h^2 k_2 f''(x) E_b(x; n, h) + E_b^2(x; n, h) \right) dx \right| \\ &\leq \left( h^2 k_2 C_2 + \frac{C_3 k_3}{6} h^3 \right) \int_{\mathbb{R}} |E_b(x; n, h)| dx \\ &\leq \left( h^2 k_2 C_2 + \frac{C_3 k_3}{6} h^3 \right) \frac{h^3}{6} I_3 k_3. \end{aligned} \quad (5.10)$$

By (H5) and (H6) we have  $\int_{\mathbb{R}} (f''(x))^2 dx < \infty$ . Hence by (5.9) and (5.10) we obtain (5.6).  $\square$

**Lemma 5.2 (variation expansion).** *Let  $K$  satisfy (H3) and (H4), with  $r = 2$ . Let  $f$  satisfy (H5) and (H6), and  $\widehat{f}_{n,h}(x)$  be the estimator of  $f$  given by (5.1). Then*

(i)  $\mathbb{V}(\widehat{f}_{n,h})$  is given by

$$[\mathbb{V}(\widehat{f}_{n,h})](x) = f(x) \frac{1}{nh} \int_{\mathbb{R}} K^2(t) dt + E_V(x; n, h), \quad x \in \mathbb{R} \quad (5.11)$$

with

$$E_V(x; n, h) = -\frac{1}{n} \left( \int_{\mathbb{R}} t K^2(t) \int_0^1 f'(x - ht u) du dt + \left( f(x) + \text{bias}(\hat{f}_{n,h})(x) \right)^2 \right) \quad (5.12)$$

(ii) The term  $E_V(x; n, h)$  satisfies

$$\begin{aligned} \mathcal{E}_V(n, h) &= \left| \int_{\mathbb{R}} E_V(x) dx \right| \\ &\leq \frac{1}{n} \left( 2M_0 + C_2 h^2 k_2 I_2 + (k_2 C_2 + \frac{C_3 k_3}{3} h) \frac{h^5}{6} I_3 k_3 \right). \end{aligned} \quad (5.13)$$

*Proof.* Using (5.2) and the fact that  $X_i, i = 1, \dots, n$ , are i.i.d. we obtain

$$\begin{aligned} \mathbb{V}(\hat{f}_{n,h}(x)) &= \mathbb{V}\left(\frac{1}{h} K\left(\frac{x - X_1}{h}\right)\right) \\ &= \frac{1}{n} \int_{\mathbb{R}} \frac{1}{h^2} K^2\left(\frac{x - y}{h}\right) f(y) dy - \frac{1}{n} \left( \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x - y}{h}\right) f(y) dy \right)^2 \\ &= \frac{1}{nh} \int_{\mathbb{R}} K^2(t) f(x - ht) dt - \frac{1}{n} \left( f(x) + \text{bias}(\hat{f}_{n,k})(x) \right)^2 \\ &= \frac{1}{nh} \int_{\mathbb{R}} K^2(t) f(x) dt + \frac{1}{nh} \int_{\mathbb{R}} K^2(t) \left( \int_x^{x-ht} f'(z) dz \right) dt \\ &\quad - \frac{1}{n} \left( f(x) + \text{bias}(\hat{f}_{n,k})(x) \right)^2 \\ &= \frac{1}{nh} \int_{\mathbb{R}} K^2(t) f(x) dt - \frac{1}{n} \int_{\mathbb{R}} t K^2(t) \int_0^1 f'(x - ht u) du dt \\ &\quad - \frac{1}{n} \left( f(x) + \text{bias}(\hat{f}_{n,k})(x) \right)^2 \end{aligned}$$

which proves (5.11) and (5.12).

We next estimate the terms

$$E_1(x) := \int_{\mathbb{R}} t K^2(t) \left( \int_0^1 f'(x - ht u) du \right) dt, \quad E_2(x) := \left( f(x) + \text{bias}(\hat{f}_{n,k})(x) \right)^2.$$

Observe that (H5)-(H6) imply

$$\int_{\mathbb{R}} |f'(x)| dx = \int_{\mathbb{R}} |f'(x + \alpha)| dx := I_1 < \infty$$

for any  $\alpha \in \mathbb{R}$ . Then using Tonelli's Theorem and (H4) we obtain

$$\begin{aligned} \int_{\mathbb{R}} |E_1(x)| dx &\leq \int_{\mathbb{R}} |t| K^2(t) \left( \int_{\mathbb{R}} \int_0^1 |f'(x - ht u)| du dx \right) dt \\ &\leq \int_{\mathbb{R}} |t| K^2(t) \left( \int_0^1 \left( \int_{\mathbb{R}} |f'(x - ht u)| dx \right) du \right) dt \leq I_1 C k_1 \end{aligned}$$

Since  $E_1$  is integrable we can use Fubini's Theorem and this yields

$$\begin{aligned} \int_{\mathbb{R}} E_1(x) dx &= \int_{\mathbb{R}} t K^2(t) \left( \int_{\mathbb{R}} \int_0^1 f'(x - ht u) du dx \right) dt \\ &= \int_{\mathbb{R}} t K^2(t) \left( \int_0^1 \left( \int_{\mathbb{R}} f'(x - ht u) dx \right) du \right) dt = 0 \end{aligned}$$

where we used the fact that  $\lim_{x \rightarrow \pm\infty} f(x) = 0$ . Next, by (H5) and (5.4) we get

$$\begin{aligned} \int_{\mathbb{R}} |E_2(x)| dx &\leq 2 \int_{\mathbb{R}} \left( f^2(x) + \text{bias}^2(\widehat{f}_{n,h})(x) \right) dx \\ &\leq 2M_0 + C_2 h^2 k_2 I_2 + \left( k_2 C_2 + \frac{C_3 k_3}{6} h \right) \frac{h^5}{3} I_3 k_3. \end{aligned}$$

Combining the above estimates we obtain (5.13).  $\square$

**Lemma 5.3 (kernel autocorrelation).** *Let  $K$  satisfy (H3) and (H4), then the function*

$$K_2(z) = \int_{\mathbb{R}} K(s) K(s-z) ds \geq 0, \quad z \in \mathbb{R}$$

*satisfies*

$$\int_{\mathbb{R}} K_2(z) dz = 1, \quad \int_{\mathbb{R}} z K_2(z) dz = 0.$$

*Moreover, for any sufficiently smooth  $f(x)$*

$$\int \frac{1}{h} K_2 \left( \frac{z-x}{h} \right) f(z) dz = f(x) + E_{C,f} \quad \text{with} \quad |E_{C,f}| \leq \|f''\|_{\infty} k_2 h^2.$$

*Proof.* Since  $K \geq 0$  we have  $K_2 \geq 0$ . Moreover, we have

$$\int_{\mathbb{R}} K_2(z) dz = \iint_{\mathbb{R} \times \mathbb{R}} K(s) K(s-z) dz ds = 1$$

and this proves the first property. Similarly, recalling that  $\int z K(z) dz = 0$ , we obtain

$$\int_{\mathbb{R}} z K_2(z) dz = \int_{\mathbb{R}} K(s) \int_{\mathbb{R}} (z-s+s) K(s-z) dz ds = 0.$$

Next, we take any smooth function  $f$  and compute

$$\begin{aligned} \int \frac{1}{h} K_2 \left( \frac{z-x}{h} \right) f(z) dz &= \int K_2(u) f(x-hu) du \\ &= f(x) + \int K_2(u) \int_x^{x-hu} f''(t)(t-x+hu) dt du. \end{aligned}$$

Finally, we estimate the last term in the above formula as follows

$$\begin{aligned} &\left| \int K_2(u) \int_x^{x-hu} f''(t)(t-x+hu) dt du \right| \\ &\leq \|f''\|_{\infty} \int K_2(u) \frac{h^2 u^2}{2} du \\ &= \frac{\|f''\|_{\infty} h^2}{2} \left( \int K(s) \int (s-u)^2 K(s-u) du ds + \int s^2 K(s) \int K(s-u) du ds \right) \\ &\leq \|f''\|_{\infty} k_2 h^2. \end{aligned}$$

$\square$



**Lemma 5.4 (product expectation).** *Let  $K$  satisfy (H3) and (H4), with  $r = 2$ . Let  $f$  be a probability density function that satisfies (H5) and (H6), and let  $\hat{f}_{n,h}(x)$  be an estimate of  $f$  given by (5.1). Then*

$$\mathbb{E}[\hat{f}_{n,h}(x)\hat{f}_{n,h}(y)] - \mathbb{E}[\hat{f}_{n,h}(x)]\mathbb{E}[\hat{f}_{n,h}(y)] = \frac{1}{Nh}f(x)K_2\left(\frac{x-y}{h}\right) - E_{\Pi}, \quad (5.14)$$

where the error term

$$E_{\Pi} = \frac{1}{N} \int \left( sK(s)K\left(s - \frac{x-y}{h}\right) \left( \int_0^1 f'(x-shu) du \right) \right) ds + \frac{1}{N} \mathbb{E}[\hat{f}(x)]\mathbb{E}[\hat{f}(y)]$$

satisfies

$$\begin{aligned} |E_{\Pi}(x, y)| &\leq \frac{C_{\Pi}}{N}, \quad \left| \int \int E_{\Pi}(x, y) dx dy \right| \leq \frac{1}{N} \left( 1 + \frac{I_3 k_3 h^3}{6} \right)^2 \\ \int \int |E_{\Pi}(x, y)| dx dy &\leq \frac{1}{N} \left( 1 + k_1 I_1 \frac{I_2 k_2 h^2}{2} + \frac{I_3 k_3 h^3}{6} \right)^2 \end{aligned} \quad (5.15)$$

for some constant  $C_{\Pi}$  and constants  $I_2, I_3$  given in (H6) and  $K_2$  defined in Lemma 5.3.

*Proof.* By the definition of the estimator  $\hat{f}$  we have

$$\mathbb{E}\left(\hat{f}(x)\hat{f}(y)\right) = \mathbb{E}\left(\frac{1}{N^2 h^2} \sum_{i,j=1}^N K\left(\frac{x-X_i}{h}\right) K\left(\frac{y-X_j}{h}\right)\right). \quad (5.16)$$

Since all  $\{X_i\}_{i=1}^N$  are i.i.d. we can split the calculation into two parts, one for the part, where the indexes coincide and the part, where indexes are different. We then can use the independence of the samples to simplify the calculation

$$\begin{aligned} \mathbb{E}\left(\hat{f}(x)\hat{f}(y)\right) &= \frac{1}{N^2 h^2} \mathbb{E}\left(\sum_{i=j} K\left(\frac{x-X_i}{h}\right) K\left(\frac{y-X_i}{h}\right)\right) \\ &\quad + \frac{1}{N^2 h^2} \mathbb{E}\left(\sum_{i \neq j} K\left(\frac{x-X_i}{h}\right) K\left(\frac{y-X_j}{h}\right)\right) \\ &= \frac{1}{Nh^2} \left[ \mathbb{E}\left(K\left(\frac{x-X}{h}\right) K\left(\frac{y-X}{h}\right)\right) \right] + \left(1 - \frac{1}{N}\right) \mathbb{E}[\hat{f}(x)]\mathbb{E}[\hat{f}(y)] \end{aligned} \quad (5.17)$$

where  $X = X_1$ . The first expectation term in (5.17) can be expanded as

$$\begin{aligned} &\frac{1}{Nh^2} \mathbb{E}\left[K\left(\frac{x-X}{h}\right) K\left(\frac{y-X}{h}\right)\right] \\ &= \frac{1}{Nh^2} \int K\left(\frac{x-t}{h}\right) K\left(\frac{y-t}{h}\right) f(t) dt \\ &= \frac{1}{Nh} \int K(s) K\left(s - \frac{x-y}{h}\right) \left(f(x) + \int_x^{x-sh} f'(z) dz\right) ds \\ &= f(x) \frac{1}{Nh} K_2\left(\frac{x-y}{h}\right) \\ &\quad - \frac{1}{N} \int sK(s) K\left(s - \frac{x-y}{h}\right) \left(\int_0^1 f'(x-shu) du\right) ds \end{aligned}$$

Let us denote

$$E_{\Pi,1} = \frac{1}{N} \int \left( sK(s) K\left(s - \frac{x-y}{h}\right) \left(\int_0^1 f'(x-shu) du\right) \right) ds, \quad E_{\Pi,2} = \frac{1}{N} \mathbb{E}[\hat{f}(x)]\mathbb{E}[\hat{f}(y)].$$

Then we obtain

$$\begin{aligned} & \mathbb{E}\left(\widehat{f}_{n,h}(x)\widehat{f}_{n,h}(y)\right) - \mathbb{E}[\widehat{f}_{n,h}(x)]\mathbb{E}[\widehat{f}_{n,h}(y)] \\ &= f(x)\frac{1}{Nh}K_2\left(\frac{x-y}{h}\right)ds - (E_{\Pi,1} + E_{\Pi,2}). \end{aligned}$$

and this establishes (5.14).

Observe that (H3), (H4) and (H5) imply

$$|E_{\Pi,1}| \leq \frac{C_1 C k_1}{N}.$$

Next, according to (5.2) and (5.4)

$$|\mathbb{E}[\widehat{f}(x)]| \leq C_0 + \frac{C_2 k_2 h^2}{2} + \frac{C_3 k_3 h^3}{6} \quad \text{for all } x \in \mathbb{R}$$

where  $C_2$  and  $C_3$  are constants from (H5) and hence

$$|E_{\Pi,2}| \leq \frac{1}{N} \left( C_0 + \frac{C_2 k_2 h^2}{2} + \frac{C_3 k_3 h^3}{6} \right)^2.$$

Combining the above estimate we conclude that

$$|E_{\Pi}| = |E_{\Pi,1} + E_{\Pi,2}| \leq \frac{1}{N} \left( C_1 C k_1 + \left( C_0 + \frac{C_2 k_2 h^2}{2} + \frac{C_3 k_3 h^3}{6} \right)^2 \right).$$

To obtain bounds on the integral of the error term, let us consider each component of the error separately. The term  $E_{\Pi,1}$  is integrable

$$\begin{aligned} \iint |E_{\Pi,1}(x, y)| dx dy &\leq \frac{1}{N} \iiint_{\mathbb{R}^3} |s| K(s) K\left(s - \frac{x-y}{h}\right) \left( \int_0^1 |f'(x - shu)| du \right) ds dx dy \\ &\leq \frac{1}{N} \int_{\mathbb{R}} |s| K(s) \left( \int_0^1 \int_{\mathbb{R}} |f'(x - shu)| dx du \right) ds \leq \frac{k_1 I_1}{N} \end{aligned} \tag{5.18}$$

Next using Fubini Theorem, we obtain

$$\begin{aligned} & \left| \iint E_{\Pi,1}(x, y) dx dy \right| \\ &\leq \frac{1}{N} \left| \iiint_{\mathbb{R}^3} s K(s) K\left(s - \frac{x-y}{h}\right) \left( \int_0^1 f'(x - shu) du \right) ds dx dy \right| \\ &= \frac{1}{N} \left| \int_{\mathbb{R}} s K(s) \left( \int_0^1 \int_{\mathbb{R}} f'(x - shu) dx du \right) ds \right| = 0. \end{aligned}$$

Therefore, using Lemma 5.1, (5.2), (5.4) and the hypothesis (H6) we obtain

$$\left| \iint_{\mathbb{R}^2} E_{\Pi}(x, y) dx dy \right| = \frac{1}{N} \left| \int_{\mathbb{R}} \mathbb{E}[\widehat{f}(x)] dx \right|^2 \leq \frac{1}{N} \left( 1 + \frac{I_3 k_3 h^3}{6} \right)^2.$$

Finally, directly from (5.18), (5.2) and (5.4) we obtained

$$\begin{aligned} \iint_{\mathbb{R}^2} |E_{\Pi}(x, y)| dx dy &\leq \iint_{\mathbb{R}^2} |E_{\Pi,1}(x, y)| dx dy + \iint_{\mathbb{R}^2} |E_{\Pi,2}(x, y)| dx dy \\ &\leq \frac{k_1 I_1}{N} + \frac{1}{N} \left( 1 + \frac{I_2 k_2 h^2}{2} + \frac{I_3 k_3 h^3}{6} \right)^2 \end{aligned}$$

□

**Theorem 5.5 (MISE expansion).** *Let  $K$  satisfy (H3) and (H4), with  $r = 2$ . Let  $f$  be a probability density function that satisfies (H5) and (H6), and let  $\hat{f}_{n,h}(x)$  be an estimate of  $f$  given by (5.1). Then*

$$\text{MISE}(\hat{f}_{n,h}) = \frac{h^4 k_2^2}{4} \int_{\mathbb{R}} (f''(x))^2 dx + \frac{1}{nh} f(x) \int_{\mathbb{R}} K^2(t) dt + \mathcal{E}_b(n, h) + \mathcal{E}_V(n, h) \quad (5.19)$$

with  $\mathcal{E}_b$  and  $\mathcal{E}_V$  defined in (5.10) and (5.13), respectively. Moreover, for every  $H > 0$  there exists  $C_{f,K,H}$  such that

$$|\mathcal{E}_b(h, n) + \mathcal{E}_V(h, n)| \leq C_{f,K,H} \left( h^5 + \frac{1}{n} \right) \quad (5.20)$$

for all  $n \geq 1$  and  $H \geq h > 0$ .

*Proof.* It is easy to show (see [25]) that

$$\begin{aligned} \text{MISE}(\hat{f}_{n,h}) &= \int_{\mathbb{R}} \mathbb{E}[\hat{f}_{n,h}(x) - f(x)]^2 dx \\ &= \int_{\mathbb{R}} (\text{bias}(\hat{f}_{n,h})(x))^2 dx + \int_{\mathbb{R}} \mathbb{V}(\hat{f}_{n,h}(x)) dx. \end{aligned}$$

and hence the result follows from Lemma 5.1 and Lemma 5.2.  $\square$

## References

- [1] N. Atkinson, An introduction in Numerical analysis. John Wiley & Sons, 1989.
- [2] D. B. H. Cline and J. D. Hart, Kernel density estimation of densities with discontinuities or discontinuous derivatives, *Statistics* (1991), 22-1, 69-84
- [3] J. E. Chacón, T. Duong, Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices, *Test* (2010), 19-2, 375–398
- [4] D. B. H. Cline, Optimal kernel density estimation of densities, *Ann. Inst. Statist. Math.* (1990), 42-2, 287-303
- [5] T. Duong, M. L. Hazelton, Cross validation bandwidth matrices for multivariate kernel density estimation, *Scandinavian Journal of Statistics* (2005).
- [6] C. van Eeden, Mean integrated squared error of kernel estimators when the density and its derivative are not necessarily continuous, *Ann. Inst. Statist. Math.* (1985), 37-A, 461-472
- [7] M. Rosenblatt, *Annals of Mathematical Statistics*, 27-3, 1956, 832-837
- [8] V.A. Epanechnikov, Non-parametric estimation of a multivariate probability density, *Theory Prob. Appl.* 14, 153-158
- [9] E. Parzen, On estimation of a probability density function and mode, *The annals of mathematical statistics*, 1065-1076, (1962).
- [10] B. van Es, On the expansion of the mean integrated squared error of a kernel density estimator, *Statistics and Probability Letters* (2000), 52, 441-450
- [11] Z. Huang, A. Gelman, Sampling for Bayesian computation with large datasets, Technical Report, (2005), Columbia University Department of Statistics.
- [12] K.B. Laskey, J.W. Myers, Population Markov chain Monte Carlo (2003), *Machine Learning*, 50, 175-196.

- [13] L.M. Le Cam, L.G. Yang, *Asymptotics in Statistics: Some Basic Concepts*, (2003), Springer-Verlag, New York.
- [14] J. Langford, A.J. Smola, M. Zinkevich, Slow learners are fast. In: Bengio Y, Schuurmans D, J.D. Lafferty, C.K.I. Williams, A. Culotta, *Advances in Neural Information Processing Systems* (2009), 22 (NIPS), New York: Curran Associates, Inc.
- [15] L.M. Murray, Distributed Markov chain Monte Carlo, in *Proceedings of Neural Information Processing Systems workshop on learning on cores, clusters and clouds.*, Volume 11.
- [16] A. Miroshnikov, E. Conlon, parallelMCMCcombine: An R Package for Bayesian Methods for Big Data and Analytics, *PLoS ONE* (2014), 9(9): e108425. DOI:10.1371/journal.pone.0108425.
- [17] A. Miroshnikov, Z. Wei, E. Conlon, Parallel Markov Chain Monte Carlo for Non-Gaussian Posterior Distributions, *Stat. Accepted* (2015).
- [18] D. Newman, A. Asuncion, P. Smyth, M. Welling, Distributed algorithms for topic models. *J Machine Learn Res* (2009), 10, 1801-1828.
- [19] W. Neiswanger, C. Wang, E.P. Xing Asymptotically Exact, Embarrassingly Parallel MCMC, *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence. 2014; pp. 623-632.*
- [20] X. Wang and D.B. Dunson, Parallelizing MCMC via Weierstrass Sampler (preprint).
- [21] E. Parzen, On estimation of a probability density function and mode, *Annals of Mathematical Statistics* (1962), 33-3, 1065-1076
- [22] H. Rue, S. Martino, N. Chopin, Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society Series B*, (2009), 71, 319-392.s
- [23] S.L. Scott, A.W. Blocker, F.V. Bonassi, Bayes and big data: The consensus Monte Carlo algorithm. *Bayes* 250 (2014).
- [24] A. Smola, S. Narayanamurthy, An architecture for parallel topic models. *Proceedings of the VLDB Endowment* (2010), 3, 1-2, 703-710.
- [25] B.W. Silverman, Density estimation for statistics and data analysis, Springer-Science+Business Media, B.V. (1986)
- [26] B.W. Simonoff, Smoothing methods in statistics, Springer (1996).
- [27] S. Zhang and R. J. Karunamuni, On kernel density estimation near endpoints, *Annals of Mathematical Statistics* (1997), *J. Stat. Plan. Infer.* (1998), 70, 301-316
- [28] D. Wilkinson, Parallel Bayesian computation. in *Kontoghiorghesm, EJ, Handbook of Parallel Computing and Statistics* (2006), Marcel Dekker/CRC Press, New York.
- [29] M.P. Wand, M.C. Jones, Multivariate plug-in bandwidth selection, *Computational Statistics* (1994).
- [30] A.W. Van der Vaart, *Asymptotic Statistics*, (1998) , Cambridge University Press, Cambridge.
- [31] S. Minsker, S. Srivastava, L. Lin, D. Dunson , Scalable and robust Bayesian inference via the median posterior, *Proceedings of the 31st International Conference on Machine Learning*, 2014, (ICML-14).

- [32] S. Srivastava, V. Cevher, Q. Tran-Dinh, D. B. Dunson, WASP: Scalable Bayes via barycenters of subset posteriors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* (2015).
- [33] M. Xu, B. Lakshminarayanan, Y. W. Teh, J. Zhu, B. Zhang, Distributed Bayesian posterior sampling via moment sharing, *Advances in Neural Information Processing Systems* (2014).